Sparse Linear Models and l1−Regularized 2SLS with High-Dimensional Endogenous Regressors and Instruments

Ying Zhu

University of California, Berkeley

20 July 2015

Sparse Linear Models and l_1-Regularized 2SLS with High-Dimensional Endogenous Regressors and Instruments*

Ying Zhu
This version (v4): August 2017
First version: August 2013

Abstract

We explore the validity of the 2-stage least squares estimator with l_1-regularization in both stages, for linear regression models where the numbers of endogenous regressors in the main equation and instruments in the first-stage equations can exceed the sample size, and the regression coefficients are sufficiently sparse. For this l_1-regularized 2-stage least squares estimator, finite-sample performance bounds are established. We then provide a simple practical method (with asymptotic guarantees) for choosing the regularization parameter. We show that this practical method can produce an l_2-consistent 2SLS estimator whose rate of convergence can be made as arbitrarily close as the scaling of our finite-sample performance bounds under quite standard conditions.

JEL Classification: C14, C31, C36
Keywords: High-dimensional statistics; Lasso; sparse linear models; endogeneity; two-stage least squares

1 Introduction

The objective of this paper is consistent estimation of regression coefficients in models with a large number of endogenous regressors and instruments. We consider the linear model

$$Y_i = X_i\beta^* + \epsilon_i = \sum_{j=1}^{p} X_{ij}\beta_j^* + \epsilon_i, \quad i = 1, ..., n \tag{1}$$

where ϵ_i is a zero-mean random error possibly correlated with X_i and β^* is a vector of unknown parameters of interest. The j^{th} component of β^* is denoted by β_j^*. The j^{th} component, X_{ij}, of the $1 \times p$ vector, X_i, is *endogenous* if $\mathbb{E}(X_{ij}\epsilon_i) \neq 0$, and *exogenous* if $\mathbb{E}(X_{ij}\epsilon_i) = 0$.

When endogenous regressors are present, the classical least squares estimator will be inconsistent for β^* (i.e., $\hat{\beta}_{OLS} \xrightarrow{p} \beta^*$) even when the dimension p of β^* is fixed and small relative to the sample size n. The two-stage least squares (2SLS) estimation plays an important role in accounting for endogeneity that comes from individual choice or market equilibrium (e.g., Wooldridge, 2010), and is based on the following "first-stage" equations for the components of X_i,

$$X_{ij} = Z_{ij}\pi_j^* + \eta_{ij} = \sum_{l=1}^{d_j} Z_{ijl}\pi_{jl}^* + \eta_{ij}, \quad i = 1, ..., n, \ j = 1, ..., p. \tag{2}$$

*I thank my PhD committee members James Powell and Martin Wainwright for useful comments. I also thank the editor Jianqing Fan, the AE, and the anonymous referees for detailed feedback and suggested improvement on this paper. I am also grateful to Jeffrey Wooldridge and Tim Vogelsang for valuable discussions. All errors are my own. This work was financially supported by Haas School of Business at U.C. Berkeley where I was a PhD student.

For each $j = 1, ..., p$, Z_{ij} is a $1 \times d_j$ vector of instrumental variables, η_{ij} a zero-mean random error which is uncorrelated with Z_{ij}, and π_j^* is a vector of unknown nuisance parameters. We will refer to the equation in (1) as the main equation and the equations in (2) as the first-stage equations. In particular, the assumption $\mathbb{E}(Z_{ij}\epsilon_i) = \mathbb{E}(Z_{ij}\eta_{ij}) = \mathbf{0}$ for all $j = 1, ..., p$ and $\mathbb{E}(Z_{ij}\eta_{ij'}) = \mathbf{0}$ for all $j \neq j'$ implies a triangular simultaneous equations model structure. For notational convenience, we will assume throughout the paper that all regressors in (1) are endogenous and $d_j = d \geq (n \vee 2)$ in (2) for all j. Our primary interest concerns the regime where $p \geq (n \vee 2)$, β^* and π_j^*s are sufficiently sparse (meaning that the ordered coefficients in β^* and π_j^* decay at sufficiently fast rates, which will be formalized in Section 2). The modification to allow $p < (n \vee 2)$ and/or $d_j \neq d_{j'}$ for $j \neq j'$ is straightforward.

Theoretical analysis for linear regression models with high dimensional endogeneity is important for applications concerning the estimation of peer effects. For example, Manresa (2015) investigates how a firm's production output is influenced by the investment of other firms. As a future extension, she suggests an alternative model that looks at the effects of peers' output rather than their investment:

$$Y_{it} = \alpha_i^* + X_{it}\theta^* + \sum_{j \in \{1,...,n\}, j \neq i} \beta_{ji}^* Y_{jt} + \epsilon_{it}, \quad i = 1, ..., n, \ t = 1, ..., T$$

where T is fixed and small relative to n; X_{it} denotes a vector of exogenous regressors specific to firm i at period t, α_i^* is the fixed effect of firm i, and β_{ji}^* is interpreted as the peer effect of firm j's output on firm i's output, where the effect of firm j on firm i is allowed to differ from the effect of firm i on firm j. Note that Y_{jt}s, the output of other firms enters the right-hand-side of the equations above as regressors and consequently, endogeneity arises from the simultaneity of the output variables when $\beta_{ji}^* \neq 0$. In this example, the number of endogenous regressors is $n(n-1)$.

For statistical models where the dimension of parameters is comparable to or even larger than the sample size, regularization methods have been given a great deal of attention (see, e.g., Bühlmann and van de Geer, 2011). Recently, these methods have been applied in a number of econometric papers. For example, Caner (2009) studies a Lasso type GMM estimator. Alternative penalized "Method of Moments" type estimators have been proposed by Gautier and Tsybakov (2014) as well as Fan and Liao (2014). Rosenbaum and Tsybakov (2010) study the high-dimensional errors-in-variables problem where the non-random regressors are observed with additive error and they present an application to hedge fund portfolio replication. Belloni, Chen, Chernozhukov, and Hansen (2012) estimate the optimal instruments using the Lasso; in an empirical example dealing with the effect of judicial eminent domain decisions on economic outcomes, they find the Lasso-based instrumental variable estimator outperforms an intuitive benchmark. Fan, Lv, and Li (2011) review the literature on sparse high-dimensional econometric models and also cover other regularization methods for the vector autoregressive model that measures the effects of monetary policy, panel data model that forecasts home price, and volatility matrix estimation in finance.

For the triangular simultaneous equations structure (1) and (2), the case where $d \geq n$, p is fixed and small relative to n, has been considered by Belloni and Chernozhukov (2011), where they show the instruments selected by the Lasso estimator in the first-stage regression can produce an efficient estimator with a small bias at the same time. In the case where $p \geq n$ and $d \geq n$, we can obtain the fitted regressors by performing a regression with the Lasso on each of the first-stage equations separately and then apply another Lasso estimation using these fitted regressors in the second stage. For convenience, we will refer to such a 2SLS estimator as the high-dimensional 2SLS (H2SLS). Despite that the H2SLS appears a natural generalization of the standard 2SLS for the case where $p \geq n$, the theoretical properties of the H2SLS have not been established in the literature.

When analyzing the H2SLS, one challenge lies in that the estimation error from each regression in the first stage accumulates in the fitted regressors $\hat{X}_j$ for all $j = 1, ..., p$; another challenge comes from the fact that the $p \times p$ random matrix $\frac{\hat{X}^T \hat{X}}{n}$ has rank at most n since $p \geq n$, where $\hat{X} = \left(\hat{X}_j\right)_{j=1}^p$ is a $n \times p$ matrix. Nevertheless, we are able to show that $\hat{v}^{0T} \frac{\hat{X}^T \hat{X}}{n} \hat{v}^0$ ($\hat{v}^0 = \hat{\beta}_{H2SLS} - \beta^*$ and $\hat{\beta}_{H2SLS}$ is our second-stage estimator) can be indeed bounded away from zero with high probability, as long as the eigenvalues of the population matrix $\mathbb{E}\left[\frac{1}{n} X^{*T} X^*\right]$ are bounded away from zero, where $X_j^* := Z_j \pi_j^*$ for $j = 1, ..., p$, $X^* = \left(X_j^*\right)_{j=1}^p$ is a $n \times p$ matrix, $Z_j = (Z_{ij})_{i=1}^n$ is a $n \times d$ matrix. This result allows $\hat{\beta}_{H2SLS}$ to achieve good finite sample (and asymptotic) properties.

We also provide a simple practical method for choosing the regularization parameter. The resulting H2SLS can be used as an initial estimator that existing inference procedures (e.g., Zhang and Zhang, 2014) can be built upon to construct confidence intervals for any coefficient in (1). Compared to the existing 2SLS techniques which limit the number of regressors entering the main equation, our H2SLS is more flexible and particularly powerful for applications where the researchers lack information about the important explanatory variables and instruments. Relative to the "Method of Moments" type estimators (such as Gautier and Tsybakov, 2014) which rely on more sophisticated optimization algorithms, the H2SLS is intuitive and can be easily implemented using built-in commands in software packages (e.g., *Stata, matlab*, or *R*) for the standard Lasso estimation of linear models without endogeneity. These features can potentially make the H2SLS very attractive to empirical researchers in economics.

We begin with a summary of notations used in this paper. The H2SLS estimator and its finite sample properties are presented in Section 2, where we also provide a practical procedure (with asymptotic guarantees) for choosing the regularization parameter. This practical procedure is tested on simulated data in Section 3. Section 4 sketches future directions of this paper. One direction regards the high dimensional "control function" approach, which is a close alternative to the H2SLS. Another direction regards inference strategies that can be built upon the H2SLS. The technical details are collected in Appendices A and B.

Notation. For the convenience of the reader, we summarize here the notations to be used throughout this paper. The letter e denotes the exponential constant. The l_q-norm of a vector $v \in m \times 1$ is denoted by $|v|_q$, $1 \leq q \leq \infty$, where $|v|_q := (\sum_{i=1}^m |v_i|^q)^{1/q}$ when $1 \leq q < \infty$ and $|v|_q := \max_{i=1,...,m} |v_i|$ when $q = \infty$. Let $J(v) = \{j \in \{1, ..., m\} \,|\, v_j \neq 0\}$ be the support of v. The cardinality of a set $J \subseteq \{1, ..., m\}$ is denoted by $|J|$. Let $|v|_0$ be the number of nonzero components in v. Given a set S, let $v_S \in m \times 1$ be the vector that has the same coordinates as v on S and zero coordinates on the complement S^c of S. For a matrix $A \in \mathbb{R}^{m \times m}$, write $|A|_\infty := \max_{i,j} |a_{ij}|$ to be the elementwise $l_\infty-$norm of A; the minimum eigenvalue of A is denoted by $\lambda_{\min}(A)$ and the maximum eigenvalue of A is denoted by $\lambda_{\max}(A)$. For functions $f(n)$ and $g(n)$, write $f(n) \gtrsim g(n)$ to mean that $f(n) \geq c g(n)$ for a universal constant $c \in (0, \infty)$ and similarly, $f(n) \precsim g(n)$ to mean that $f(n) \leq c' g(n)$ for a universal constant $c' \in (0, \infty)$; $f(n) \asymp g(n)$ when $f(n) \gtrsim g(n)$ and $f(n) \precsim g(n)$ hold simultaneously. Denote $\max\{a, b\}$ by $a \vee b$ and $\min\{a, b\}$ by $a \wedge b$. As a general rule for this paper, c constants denote positive universal constants that are independent of n.

2 High-dimensional 2SLS estimation

For the first-stage estimation, we consider

$$\hat{\pi}_j \in \text{argmin}_{\pi_j \in \mathbb{R}^d} \frac{1}{2n} |X_j - Z_j \pi_j|_2^2 + \lambda_{n,j} |\pi_j|_1 \tag{3}$$

for $j = 1, ..., p$. Denote the fitted regressors using the first-stage estimates by $\hat{X}_j := Z_j\hat{\pi}_j$ for $j = 1, ..., p$, and $\hat{X} = \left(\hat{X}_j\right)_{j=1}^p$. For the second-stage estimation, we consider

$$\hat{\beta}_{H2SLS} \in \text{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{2n}|Y - \hat{X}\beta|_2^2 + \lambda_n |\beta|_1 . \tag{4}$$

Remark. After (3), an extra step, which performs an OLS with the regressors selected by $\hat{\pi}_j$ to obtain $\hat{\pi}_j^{OLS}$ for $j = 1, ..., p$, may be used before (4). In the third step, we apply the Lasso to estimate the main equation parameters with the fitted regressors based on $\hat{\pi}_j^{OLS}$s. This type of procedure is analogous to those in Candès and Tao (2007), Belloni and Chernozhukov (2013), for example.

In the literature on the Lasso estimation of $Y_i = X_i\beta^* + \epsilon_i$ with exogenous X_i, one typically assumes (or shows) that $\max_j \frac{1}{n}\sum_{i=1}^n X_{ij}^2$ can be bounded from above with high probability so that X_{ij}s can be normalized to make $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1$ for all $j = 1, ..., p$ (e.g., Bickel, et. al, 2009). Similarly, in this paper, we show (in Lemma A.2) that, with high probability,

$$\max_{j,l} \frac{1}{n}\sum_{i=1}^n Z_{ijl}^2 \leq \max_{j,l} \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n Z_{ijl}^2\right) + 8e\sqrt{\frac{\log(pd)}{n}}, \tag{5}$$

$$\max_{j} \frac{1}{n}\sum_{i=1}^n \hat{X}_{ij}^2 \leq \max_{j} \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_{ij}^{*2}\right) + 4\max_{j} \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_{ij}^{*2}\right)\mathcal{T}_1, \tag{6}$$

where $\mathcal{T}_1$ is to be defined in Assumption 2.4. As a result, if $\sqrt{\frac{\log(p\vee d)}{n}}$ is of the same order as $\max_{j,l}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n Z_{ijl}^2\right)$ and $\mathcal{T}_1$ is of order 1, then $\max_{j,l}\frac{1}{n}\sum_{i=1}^n Z_{ijl}^2 \precsim \max_{j,l}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n Z_{ijl}^2\right)$ and $\max_j \frac{1}{n}\sum_{i=1}^n \hat{X}_{ij}^2 \precsim \max_j \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_{ij}^{*2}\right)$ with high probability. Without loss of generality, we will assume that $\hat{X}_{ij}$s are normalized so that $\frac{1}{n}\sum_{i=1}^n \hat{X}_{ij}^2 = 1$ for all $j = 1, ..., p$. In interpretating the final results, one needs to scale back the estimates of β^* by the normalizing factor. On a related note, we point out that the results in this paper do not depend on whether Z_{jl}s are normalized or not since our analysis relies on $\hat{\pi}_j$ only through $\sqrt{\frac{1}{n}\sum_{i=1}^n \left(Z_{ij}\hat{\pi}_j - Z_{ij}\pi_j^*\right)^2}$ and $|\hat{\pi}_j - \pi_j^*|_1 |\frac{1}{n}\sum_{i=1}^n Z_{ij}^T\eta_{ij}|_\infty$.

We begin with the finite sample analysis of $\hat{\beta}_{H2SLS}$. Guided by the finite sample bounds, we show the asymptotic behavior of $\hat{\beta}_{H2SLS}$ along with the requirement on the size of λ_n. We then develop an implementable algorithm for choosing λ_n with asymptotic guarantees.

2.1 Finite sample bounds

The first result (Theorem 2.1) requires the following assumptions.

Assumption 2.1. *The draws $(\epsilon_i, \eta_i, Z_i)_{i=1}^n$ are independently distributed, where $\eta_i = (\eta_{ij})_{j=1}^p$ and $Z_i = (Z_{ij})_{j=1}^p$; for all $j = 1, ..., p$, $\mathbb{E}(\frac{1}{n}Z_j^T\epsilon) = \mathbb{E}(\frac{1}{n}Z_j^T\eta_j) = \mathbf{0}$ and $\mathbb{E}(\frac{1}{n}Z_j^T\eta_{j'}) = \mathbf{0}$ for all $j \neq j'$.*

For a random variable V, as in Vershynin (2012), we define the "sub-Gaussian" norm $|V|_\Psi := \sup_{r\geq 1} r^{-\frac{1}{2}}\left(\mathbb{E}|V|^r\right)^{\frac{1}{r}}$.

Assumption 2.2. *For all $i = 1, ..., n$,*

(i) there exist parameters ρ_η, ρ_ϵ, ρ_Z, and ρ_{X^*} such that $\max_{j=1,...,p}|\eta_{ij}|_\Psi \leq \rho_\eta$, $|\epsilon_i|_\Psi \leq \rho_\epsilon$, $\max_{j=1,...,p,\, l=1,...,d}|Z_{ijl}|_\Psi \leq 1$, and $\max_{j=1,...,p}\left|X_{ij}^*\right|_\Psi \leq \rho_{X^*}$;

(ii) in terms of $Z_j \in \mathbb{R}^{n\times d}$, for any unit vector $a \in \mathbb{R}^d$, there exists a parameter $\tilde{\rho}_Z$ such that $\max_{j=1,...,p} \left|a^T Z_{ij}^T\right|_\Psi \leq \tilde{\rho}_Z$, where Z_{ij} is the ith row of Z_j;

(iii) in terms of $X^* \in \mathbb{R}^{n\times p}$, for any unit vector $a \in \mathbb{R}^p$, there exists a parameter $\tilde{\rho}_{X^*}$ such that $\left|a^T X_i^{*T}\right|_\Psi \leq \tilde{\rho}_{X^*}$, where X_i^* is the ith row of X^*.

Assumption 2.2 is known as the sub-Gaussian tail condition defined in Vershynin (2012). Sub-Gaussian variables constitute a reasonably general family of distributions that include Gaussian mixtures and distributions with bounded support. Assumption 2.2(i) implies that η_{ij}s, ϵ_is, Z_{ijl}s and X_{ij}^*s are sub-Gaussian variables and is used in deriving the lower bound on the regularization parameters. Note that the sub-Gaussian parameter associated with Z_{ijl}s is assumed to be 1. This assumption is only intended to lighten the notations and can be easily relaxed to a more general value, say, ρ_Z. Assumption 2.2(ii)-(iii) imply that Z_js and X^* are sub-Gaussian matrices and are only used to establish eigenvalue conditions on $\hat{v}^{jT}\frac{Z_j^T Z_j}{n}\hat{v}^j$s and $\hat{v}^{0T}\frac{X^{*T}X^*}{n}\hat{v}^0$ (where $\hat{v}^j = \hat{\pi}_j - \pi_j^*$ and $\hat{v}^0 = \hat{\beta}_{H2SLS} - \beta^*$). Assumptions like 2.2 are common in the literature on high dimensional statistics (see, e.g., Loh and Wainwright, 2012; Negahban, et. al 2012; Rosenbaum and Tsybakov, 2013).

Assumption 2.3. *$\underline{\kappa}_2 = \lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n}X^{*T}X^*\right]\right)$ is bounded away from zero; moreover, there exist a positive universal constant c^* such that*

$$\left|\Delta^T \frac{X^{*T}X^*}{n}\Delta\right| \geq \frac{\underline{\kappa}_2}{2}|\Delta|_2^2 - c^*\underline{\kappa}_2\left(\frac{\tilde{\rho}_{X^*}^4}{\underline{\kappa}_2^2}\vee 1\right)\frac{\log p}{n}|\Delta|_1^2 \qquad \forall \Delta \in \mathbb{R}^p$$

with probability at least $1 - 2\exp(-\log p)$.

Remark. The bound in Assumption 2.3 can be derived under lower level conditions (see Lemma B.2, which is a consequence of Lemmas 12, 13 and 15 in Loh and Wainwright, 2012).

To state the following assumption, we define a thresholded subset

$$S_{\tau_j} := \left\{l \in \{1, 2, ..., d\} : \left|\pi_{jl}^*\right| > \tau_j\right\} \tag{7}$$

and $k_1 = \max_{j=1,...,p}\left|S_{\tau_j}\right|$. We use $S_{\tau_j}^c$ to denote the complement of S_{τ_j}.

Assumption 2.4. *There exist positive universal constants $c^\star$, $c^\dagger$, $c^{'}$, c_0, c_1, and c_2 such that for $\lambda_{n,j} \geq c^\star\rho_\eta\sqrt{\frac{\log(p\vee d)}{n}}$ (uniformly in $j = 1, ..., p$) in (3),*

$$\max_{j=1,...,p}\left|\hat{\pi}_j - \pi_j^*\right|_2 \leq c^\dagger(err_e + err_a) \tag{8}$$

$$\max_{j=1,...,p}\left|\hat{\pi}_j - \pi_j^*\right|_1 \leq c^{'}\left(\sqrt{k_1}err_e + \sqrt{k_1}err_a + \max_{j=1,...,p}|\pi_{j,S_{\tau_j}^c}^*|_1\right) := \tilde{\mathcal{T}}_1 \tag{9}$$

$$\max_{j=1,...,p}\sqrt{\frac{1}{n}\sum_{i=1}^n\left(Z_{ij}\hat{\pi}_j - Z_{ij}\pi_j^*\right)^2} \leq c_0\bar{\kappa}_1^{\frac{1}{2}}(err_e + err_a) := \mathcal{T}_1 \tag{10}$$

with probability at least $1 - c_1\exp(-c_2\log(p\vee d))$, where $\tau_j := \underline{\kappa}_1^{-1}\lambda_{n,j}$, $err_e := \frac{\sqrt{k_1}}{\underline{\kappa}_1}\max_j\lambda_{n,j}$, $err_a := \max_j|\pi_{j,S_{\tau_j}^c}^*|_1^{\frac{1}{2}}\left(\frac{\lambda_{n,j}}{\underline{\kappa}_1}\right)^{\frac{1}{2}}$, $\underline{\kappa}_1 := \min_j\lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n}Z_j^T Z_j\right]\right)$, and $\bar{\kappa}_1 := \max_j\lambda_{\max}\left(\mathbb{E}\left[\frac{1}{n}Z_j^T Z_j\right]\right)$.

Moreover, $\underline{\kappa}_1$ is bounded away from zero and $\bar{\kappa}_1$ is bounded from above.

Assumption 2.4 imposes finite sample bounds on the first-stage estimates $\hat{\pi}_j$s. More specific forms of bounds (8)-(10) can be derived under lower level conditions; see Lemma B.3. Note that the bound in (8) consists of an estimation error (denoted by err_e) and an approximation error (denoted by err_a). The quantity err_e has the typical scaling achieved by $\left|\hat{\pi}_j - \pi^*_{j,S_{\tau_j}}\right|_2$ where $\pi^*_{j,S_{\tau_j}}$ has the same coordinates as π^*_j on S_{τ_j} and zero coordinates on the complement $S^c_{\tau_j}$ of S_{τ_j}. The quantity err_a accounts for the remaining error from $\pi^*_{j,S^c_{\tau_j}}$.

The following assumption imposes growth conditions on n, d, p, $k_1 = \max_{j=1,...,p}\left|S_{\tau_j}\right|$, and $\max_{j=1,...,p}|\pi^*_{j,S^c_{\tau_j}}|_1$.

Assumption 2.5. *In terms of ρ_{X^*} and ρ_ϵ defined in Assumption 2.2 as well as $\tilde{\mathcal{T}}_1$ defined in (9) and $\mathcal{T}_1$ defined in (10),*

(i) $\sqrt{\frac{\log(dp^2)}{n}} \leq \frac{2}{5}$, $8e\rho^2_{X^*}\sqrt{\frac{\log p}{n}} \leq \max_j \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X^{*2}_{ij}\right)$, and $8e\rho^2_\epsilon\sqrt{\frac{\log p}{n}} \leq \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \epsilon^2_i\right)$, where e is the exponential constant;

(ii) $\mathcal{T}_1 \leq \max_j \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X^{*2}_{ij}\right)$;

(iii) there exists a positive universal constant $c_0^{'}$ such that

$$8e\sqrt{\frac{\rho^2_\eta \log(dp^2)}{n}}\tilde{\mathcal{T}}_1 \leq c_0^{'}\max_j \sqrt{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X^{*2}_{ij}\right)}\mathcal{T}_1.$$

For stating Theorem 2.1, we define

$$\mathcal{T}_0 = \check{c}\max\left\{|\beta^*|_1\sigma_{X^*}\mathcal{T}_1,\ \sigma_\epsilon\mathcal{T}_1,\ \rho_{X^*}\rho_\eta|\beta^*|_1\sqrt{\frac{\log p}{n}},\ \rho_{X^*}\rho_\epsilon\sqrt{\frac{\log p}{n}}\right\} \tag{11}$$

where $\check{c}$ is some positive universal constant, $\mathcal{T}_1$ is defined in (10), ρ_{X^*}, ρ_η, ρ_ϵ are defined in Assumption 2.2, $\sigma_{X^*} := \max_j \sqrt{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X^{*2}_{ij}\right)}$, and $\sigma_\epsilon := \sqrt{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \epsilon^2_i\right)}$. As for π^*_js, we introduce a thresholded subset for β^*:

$$S_{\underline{\tau}} := \left\{j \in \{1,\ 2,\ ...,\ p\} \ :\ \left|\beta^*_j\right| > \tau\right\} \tag{12}$$

and $k_2 = |S_\tau|$. We use S^c_τ to denote the complement of S_τ.

Theorem 2.1 (Finite sample bounds). *Let λ_n in (4) satisfy $\lambda_n \geq \mathcal{T}_0$ with $\mathcal{T}_0$ defined in (11). Suppose Assumptions 2.1-2.5 hold. If*

$$|\beta^*|_1\lambda_n^{-1}\left(\frac{b_0\log p}{n} \vee \mathcal{T}_1^2\right) \leq c^{''} \qquad \text{where } b_0 = \underline{\kappa}_2\left(\frac{\tilde{\rho}^4_{X^*}}{\underline{\kappa}^2_2} \vee 1\right) \tag{13}$$

for some positive universal constant $c^{''}$, then for $\tau = \frac{\lambda_n}{\underline{\kappa}_2}$ in (12), we have

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \ \leq\ c_0^*\left(\underline{\kappa}_2^{-1}\sqrt{k_2}\lambda_n + \sqrt{\underline{\kappa}_2^{-1}|\beta^*_{S^c_\tau}|_1\lambda_n}\right) := \bar{B}, \tag{14}$$

$$|\hat{\beta}_{H2SLS} - \beta^*|_1 \ \leq\ 4\left(\sqrt{k_2}\bar{B} + |\beta^*_{S^c_\tau}|_1\right), \tag{15}$$

with probability at least $1 - c_1^* \exp(-c_2^* \log p)$*, where* c_0^**,* c_1^* *and* c_2^* *are some positive universal constants.*

The proof for Theorem 2.1 is provided in Section A.1. Under condition (13) and Assumption 2.3, we show in Lemma A.1 that $\hat{v}^{0T} \frac{\hat{X}^T \hat{X}}{n} \hat{v}^0$ (where $\hat{v}^0 = \hat{\beta}_{H2SLS} - \beta^*$) is bounded away from zero with high probability. This result allows $|\hat{\beta}_{H2SLS} - \beta^*|_2$ to achieve the bound in (14). As the bound on $\left|\hat{\pi}_j - \pi_j^*\right|_2$, the bound $\bar{B}$ on $|\hat{\beta}_{H2SLS} - \beta^*|_2$ also consists of an estimation error (which is of order $\frac{\sqrt{k_2}}{\underline{\kappa}_2}\lambda_n$) and an approximation error (which is of order $\sqrt{\frac{|\beta^*_{S_\tau^c}|_1}{\underline{\kappa}_2}\lambda_n}$); moreover, $\frac{\sqrt{k_2}}{\underline{\kappa}_2}\lambda_n$ and $\sqrt{\frac{|\beta^*_{S_\tau^c}|_1}{\underline{\kappa}_2}\lambda_n}$ have similar interpretations as err_e and err_a, respectively (see the discussion following Assumption 2.4).

From Theorem 2.1, we see that when λ_n is of the same order as $\mathcal{T}_0$, the bound on $|\hat{\beta}_{H2SLS} - \beta^*|_2$ involves $\mathcal{T}_1$ defined in (10), which gives an upper bound for the square root of the prediction errors associated with the first-stage estimates $\hat{\pi}_j$s. There are special cases where we can pin down the choice of the universal constant c_0 in $\mathcal{T}_1$; as an example, suppose we assume for all $j = 1, ..., p$:

(1) π_j^* is exactly sparse with at most k_1 non-zero components,

(2) Z_j is fixed and normalized so that $\sqrt{\frac{1}{n}\sum_{i=1}^n Z_{ijl}^2} \leq 1$ for all $l = 1, ..., d$,

(3) each fixed Z_j satisfies $\frac{|Z_j\tilde{\Delta}|_2^2}{n} \geq \underline{\kappa}_1^{RE} > 0$ and $\frac{|Z_j\tilde{\Delta}|_2^2}{n} \leq \bar{\kappa}_1^{RE} \leq \infty$ for all nonzero

$$\tilde{\Delta} \in \left\{\Delta \in \mathbb{R}^d \,:\, |\Delta_{S_{\tau_j}^c}|_1 \leq 3|\Delta_{S_{\tau_j}}|_1\right\}.$$

Then, in view of Corollary 2 in Negahban, et. al (2012), we have

$$\sqrt{\frac{1}{n}\sum_{i=1}^n \left[Z_{ij}\hat{\pi}_j - Z_{ij}\pi_j^*\right]^2} \leq 2\frac{\sqrt{\bar{\kappa}_1^{RE}}}{\underline{\kappa}_1^{RE}}\sqrt{k_1}\max_j \lambda_{n,j}. \tag{16}$$

In our context, it makes more sense that we should account for the randomness in Z_js; hence, instead of treating Z_j as fixed and working with Item (3) in the above, we impose assumptions on $\underline{\kappa}_1 := \min_{j=1,...,p} \lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n}Z_j^T Z_j\right]\right)$ and $\bar{\kappa}_1 := \max_{j=1,...,p} \lambda_{\max}\left(\mathbb{E}\left[\frac{1}{n}Z_j^T Z_j\right]\right)$ while only requiring $\mathbb{E}\left(\frac{1}{n}Z_{jl}^T\eta_j\right) = 0$ for all $j = 1, ..., p$ and $l = 1, ..., d$. This approach along with the generality of our assumption on π_j^*s (where we do not assume the exact sparsity) makes deriving a sharp choice of the universal constant c_0 in $\mathcal{T}_1$ highly difficult.

Generally speaking, the specification of universal constants in finite sample analysis is often coarse except in very simple models. Even if sharp universal constants can be obtained, the presence of unknown nuisance parameters ρ_η, $\underline{\kappa}_1$, $\bar{\kappa}_1$, k_1 and $\max_j |\pi^*_{j,S^c_{\tau_j}}|_1$ in $\mathcal{T}_1$, (10), or $\bar{\kappa}_1^{RE}$ and $\underline{\kappa}_1^{RE}$ in (16) makes setting λ_n to its optimal value nearly infeasible. In contrast, the asymptotic rates implied by the finite sample bounds are often more useful from a practical view point. For this reason, we present the following corollary which exhibits the asymptotic behavior of $\hat{\beta}_{H2SLS}$ along with the requirement on the size of λ_n. This result follows immediately from Theorem 2.1.

Corollary 2.1 (Asymptotic bounds)*. Let the conditions in Theorem 2.1 hold. Suppose*

$$\underline{\kappa}_1^{-1},\, \bar{\kappa}_1,\, \rho_\eta,\, \rho_\epsilon,\, \rho_{X^*} \;=\; O(1), \tag{17}$$

$$\max_{j=1,...,p} |\pi^*_{j,S^c_{\tau_j}}|_1 \;=\; O\left((k_1 \vee 1)\sqrt{\frac{\log(d \vee p)}{n}}\right), \tag{18}$$

and the regularization parameters satisfy

$$\sqrt{\frac{\log(d \vee p)}{n}} = O(\lambda_{n,j}) \qquad \forall\ j = 1, ..., p, \tag{19}$$

$$(|\beta^*|_1 \vee 1)\sqrt{\frac{(k_1 \vee 1)\log(d \vee p)}{n}} = O(\lambda_n). \tag{20}$$

Then as $n \to \infty$, $d \to \infty$, and $p \to \infty$, we have

$$\begin{aligned} |\hat{\beta}_{H2SLS} - \beta^*|_2 &= O_p\left(\underline{\kappa}_2^{-1}\sqrt{k_2}\lambda_n + \sqrt{\underline{\kappa}_2^{-1}|\beta^*_{S_\tau^c}|_1\lambda_n}\right), \\ |\hat{\beta}_{H2SLS} - \beta^*|_1 &= O_p\left(\underline{\kappa}_2^{-1}k_2\lambda_n + \sqrt{\underline{\kappa}_2^{-1}k_2|\beta^*_{S_\tau^c}|_1\lambda_n} + |\beta^*_{S_\tau^c}|_1\right). \end{aligned}$$

A condition like (18), which ensures the "small" coefficients decay sufficiently fast, is often assumed in the literature on approximately sparse models. Under (18), we have $\max_j \left|\hat{\pi}_j - \pi_j^*\right|_2 = O_p\left(\sqrt{\frac{(k_1 \vee 1)\log(d \vee p)}{n}}\right)$. When $k_1 > 0$, (18) corresponds to the foremost scenario where the first-stage approximation error $err_a = O\left(\sqrt{\frac{k_1 \log(d \vee p)}{n}}\right)$ in $\mathcal{T}_1$ does not dominate the first-stage estimation error err_e, which is of order $\sqrt{\frac{k_1 \log(d \vee p)}{n}}$.

Based on (20), we provide an implementable algorithm for choosing λ_n along with asymptotic guarantees in the following.

2.2 Choosing the regularization parameter

Note that the choice of λ_n in (20) depends on $|\beta^*|_1$, which is due to the fact that the second-stage procedure (4) uses the first-stage estimates $\hat{X}_j = Z_j\hat{\pi}_j$ as the surrogate of the unknown $X_j^* = Z_j\pi_j^*$. Other surrogate-type Lasso estimators such as the one in Rosenbaum and Tsybakov (2013) also involve the factor $|\beta^*|_1$. Here we propose a simple implementable algorithm for choosing λ_n, which consists of two steps: By over-penalizing, the first step uses a regularization parameter $\lambda_n = \lambda_n^{(0)}$ such that $\mathcal{T}_0 = o(\lambda_n^{(0)})$ and this $\lambda_n^{(0)}$ returns an initial estimator, $\hat{\beta}^{(1)}$, which satisfies $\left|\hat{\beta}^{(1)}\right|_1 = |\beta^*|_1 + o(1)$ with probability $1 - o(1)$; the second step tunes the amount of regularization and possibly decreases (but never increases) the rate of convergence using the initial estimator returned by Step 1. The algorithm is described below.

The main algorithm

1. *(Over-Penalization) Let $\left|\hat{\beta}^{(0)}\right|_1 = \left(\frac{n}{\log(d \vee p)}\right)^{\frac{1}{4}}$ and $\hat{k}_1 = \max_{j=1,...,p}|J(\hat{\pi}_j)|$. For any arbitrarily small number $\varsigma \in \left(0, \frac{1}{4}\right)$, form $\hat{\mathcal{T}}_1 = \sqrt{\hat{k}_1 \vee 1}\left(\frac{\log(d \vee p)}{n}\right)^{\frac{1}{2}-\varsigma}$ and perform (4) with*

$$\lambda_n = \lambda_n^{(0)} = \hat{\mathcal{T}}_0^{(0)} = \left|\hat{\beta}^{(0)}\right|_1 \hat{\mathcal{T}}_1$$

to obtain the initial estimates $\hat{\beta}^{(1)}$.

2. *(Adjusted-Penalization) For some constant $C > 0$ and the same ς as in the "Over-Penalization" step, perform (4) with*

$$\lambda_n = \lambda_n^{(1)} = \hat{\mathcal{T}}_0^{(1)} = C\left(\left|\hat{\beta}^{(1)}\right|_1 \vee 1\right)\hat{\mathcal{T}}_1 \tag{21}$$

to obtain the estimates $\hat{\beta}^{(2)}$.

Using $\left|\hat{\beta}^{(2)}\right|_1$ returned by Step 2, we can apply additional adjustment to $\lambda_n^{(1)}$ by replacing $\left|\hat{\beta}^{(1)}\right|_1$ with $\left|\hat{\beta}^{(2)}\right|_1$. Asymptotically, further iterations yield the same rate of convergence as $\hat{\beta}^{(2)}$ but may perform better within small samples. Similarly, while the choice of the constant, C, in (21) does not affect the asymptotic validity of our algorithm, it could affect the small sample performance. In practice, selecting C can be assisted with the most popular "Cross-Validation" (CV) criterion or the "Estimation-Stability-Cross-Validation" (ESCV) criterion recently proposed by Lim and Yu (2013). According to Lim and Yu (2013) as well as Yu (2013), the ESCV criterion yields a smaller-size model but similar performance in prediction relative to the CV criterion. The details on how to tailor the ESCV criterion to our "Adjusted-Penalization" step are deferred to Section 4.

The asymptotic validity of the algorithm is given by Theorem 2.2, for which we impose an additional assumption.

Assumption 2.6. $\left(\hat{k}_1 \vee 1\right) \asymp (k_1 \vee 1)$ *with probability* $1 - o(1)$.

Remark. Assumption 2.6 can be shown under lower level conditions; see Lemma B.4. Under Assumption 2.6, we have $\hat{\mathcal{T}}_1 = \sqrt{\hat{k}_1 \vee 1}\left(\frac{\log(d\vee p)}{n}\right)^{\frac{1}{2}-\varsigma} \asymp \sqrt{k_1 \vee 1}\left(\frac{\log(d\vee p)}{n}\right)^{\frac{1}{2}-\varsigma}$ with probability $1 - o(1)$.

Theorem 2.2. *Suppose* $\frac{\log(d\vee p)}{n} = o(1)$ *and* $|\beta^*|_1 = O\left(\left(\frac{n}{\log(d\vee p)}\right)^{\frac{1}{4}}\right)$. *Let Assumption 2.6, the conditions in Theorem 2.1, and (17)-(19) hold. Then, as* $n \to \infty$, $d \to \infty$, *and* $p \to \infty$,

$$\left|\hat{\beta}^{(1)} - \beta^*\right|_2 = O_p\left(\bar{B}^{(1)}\right), \tag{22}$$

$$\left|\hat{\beta}^{(1)} - \beta^*\right|_1 = O_p\left(\sqrt{k_2}\bar{B}^{(1)} + |\beta^*_{S_\tau^c}|_1\right), \tag{23}$$

where $\bar{B}^{(1)} := \frac{\sqrt{k_2}}{\underline{\kappa}_2}\mathcal{T}_0^{(0)} + \sqrt{\frac{\mathcal{T}_0^{(0)}|\beta^*_{S_\tau^c}|_1}{\underline{\kappa}_2}}$, $\mathcal{T}_0^{(0)} = \sqrt{k_1 \vee 1}\left(\frac{\log(d\vee p)}{n}\right)^{\frac{1}{4}-\varsigma}$, *and* $\hat{\beta}^{(1)}$ *are the initial estimates returned by Step 1 of the algorithm based on* $\left|\hat{\beta}^{(0)}\right|_1$. *Moreover, if* $\sqrt{k_2}\bar{B}^{(1)} + |\beta^*_{S_\tau^c}|_1 = o(1)$, *then* $\left|\hat{\beta}^{(1)}\right|_1 = |\beta^*|_1 + o(1)$ *with probability* $1 - o(1)$*; also,*

$$\left|\hat{\beta}^{(2)} - \beta^*\right|_2 = O_p\left(\bar{B}^{(2)}\right), \tag{24}$$

$$\left|\hat{\beta}^{(2)} - \beta^*\right|_1 = O_p\left(\sqrt{k_2}\bar{B}^{(2)} + |\beta^*_{S_\tau^c}|_1\right), \tag{25}$$

where $\bar{B}^{(2)} := \frac{\sqrt{k_2}}{\underline{\kappa}_2}\mathcal{T}_0^{(1)} + \sqrt{\frac{\mathcal{T}_0^{(1)}|\beta^*_{S_\tau^c}|_1}{\underline{\kappa}_2}}$, $\mathcal{T}_0^{(1)} = (|\beta^*|_1 \vee 1)\sqrt{k_1 \vee 1}\left(\frac{\log(d\vee p)}{n}\right)^{\frac{1}{2}-\varsigma}$, *and* $\hat{\beta}^{(2)}$ *are the estimates returned by Step 2 of the algorithm based on* $\left|\hat{\beta}^{(1)}\right|_1$.

The proof for Theorem 2.2 is provided in Section A.2. Note that, if $\bar{B}^{(2)} \to 0$ as $n \to \infty$, then $\hat{\beta}^{(2)}$ is l_2-consistent for β^*. Furthermore, if $\lambda_n \asymp \mathcal{T}_0$ and $1 = O\left(|\beta^*|_1\right)$ in Theorem 2.1, the rates in (24) and (25) can be made arbitrarily close to the scaling of (14) and (15), respectively.

As long as ρ_ϵ, $\rho_\eta = O(1)$ for any sub-Gaussian noise ϵ and η_js in our model, the algorithm above is asymptotically valid even though it does not account for the effects of the noise. On the other hand, the noise factors could affect the small sample performance of the H2SLS especially

when they are relatively large. In the following, we will focus on the most studied Gaussian-noise case where $\eta_{ij} \overset{i.i.d.}{\sim} \mathcal{N}\left(0,\, \sigma_\eta^2\right)$ for all $j = 1, ..., p$ and $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0,\, \sigma_\epsilon^2\right)$. Throughout the rest, we will assume $1 = O\left(\min\left(\sigma_\eta,\, \sigma_\epsilon,\, |\beta^*|_1\right)\right)$ (i.e., the noise variances and $|\beta^*|_1$ are bounded away from zero); note that this condition is only intended for lightening the notations and can be easily relaxed. In the context of Gaussian noise, ρ_η (and ρ_ϵ) only differs from σ_η (respectively, σ_ϵ) by a constant multiplier; moreover, if $1 = O\left(\sigma_\eta\right)$, condition (18) holds, and $\underline{\kappa}_1^{-1}$, $\bar{\kappa}_1 = O(1)$, we have $\mathcal{T}_1 = O\left(\left(\sqrt{\sigma_\eta^2 k_1} \vee \sqrt{\sigma_\eta}\right)\sqrt{\frac{\log(d\vee p)}{n}}\right)$. These facts motivate us to consider the *modified* algorithm as below.

The modified algorithm for i.i.d. Gaussian noise

1. *(Over-Penalization) Let* $\left|\hat{\beta}^{(0)}\right|_1 = \hat{\sigma}_\epsilon^{(0)} = \left(\frac{n}{\log(d\vee p)}\right)^{\frac{1}{4}}$, $\hat{\sigma}_\eta = \max_j \sqrt{\frac{1}{n}\sum_{i=1}^n \left(X_{ij} - Z_{ij}\hat{\pi}_j\right)^2}$, *and* $\hat{k}_1 = \max_j |J(\hat{\pi}_j)|$. *For any arbitrarily small number* $\varsigma \in \left(0,\, \frac{1}{4}\right)$, *form*

$$\hat{\mathcal{T}}_1 = \left(\sqrt{\hat{\sigma}_\eta^2 \hat{k}_1} \vee \sqrt{\hat{\sigma}_\eta}\right)\left(\frac{\log(d\vee p)}{n}\right)^{\frac{1}{2}-\varsigma}$$

and perform (4) with

$$\lambda_n = \lambda_n^{(0)} = \hat{\mathcal{T}}_0^{(0)} = \left(\frac{n}{\log(d\vee p)}\right)^{\frac{1}{4}} \max\left\{\hat{\mathcal{T}}_1,\, \hat{\sigma}_\eta\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\varsigma},\, \left(\frac{\log p}{n}\right)^{\frac{1}{2}-\varsigma}\right\} \tag{26}$$

to obtain the initial estimates $\hat{\beta}^{(1)}$.

2. *(Adjusted-Penalization) Using* $\hat{\beta}^{(1)}$ *from the "Over-Penalization" step, we form*

$$\hat{\sigma}_\epsilon^{(1)} = \sqrt{\frac{1}{n}\sum_{i=1}^n \left(Y_i - X_i\hat{\beta}^{(1)}\right)^2}. \tag{27}$$

For some constant $C > 0$ *and the same* ς *as in the "Over-Penalization" step, perform (4) with*

$$\lambda_n = \lambda_n^{(1)} = \hat{\mathcal{T}}_0^{(1)} = C\max\left\{\left(\left|\hat{\beta}^{(1)}\right|_1 \vee \hat{\sigma}_\epsilon^{(1)}\right)\hat{\mathcal{T}}_1,\, \hat{\sigma}_\eta\left|\hat{\beta}^{(1)}\right|_1\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\varsigma},\, \hat{\sigma}_\epsilon^{(1)}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\varsigma}\right\} \tag{28}$$

to obtain the estimates $\hat{\beta}^{(2)}$.

For the first-stage regularization parameters in (3), $\lambda_{n,j}$s, a simpler version of the *modified* algorithm above can be used. In the over-penalization step, we set $\hat{\sigma}_\eta^{(0)} = \left(\frac{n}{\log(d\vee p)}\right)^{\frac{1}{4}}$ and

$$\lambda_{n,j} = \lambda_{n,j}^{(0)} = \hat{\sigma}_\eta^{(0)}\left(\frac{\log\left(p\vee d\right)}{n}\right)^{\frac{1}{2}-\varsigma} \tag{29}$$

to obtain the initial estimates $\hat{\pi}_j^{(1)}$s. We then set

$$\hat{\sigma}_\eta^{(1)} = \max_{j=1,...,p}\sqrt{\frac{1}{n}\sum_{i=1}^n \left(X_{ij} - Z_{ij}\hat{\pi}_j^{(1)}\right)^2},$$

$$\lambda_{n,j} = \lambda_{n,j}^{(1)} = \hat{\sigma}_\eta^{(1)}\left(\frac{\log\left(p\vee d\right)}{n}\right)^{\frac{1}{2}-\varsigma}, \tag{30}$$

to obtain the estimates $\hat{\pi}_j^{(2)}$s, which are used to construct

$$\hat{\sigma}_\eta := \hat{\sigma}_\eta^{(2)} = \max_{j=1,\ldots,p} \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{ij} - Z_{ij}\hat{\pi}_j^{(2)}\right)^2}. \tag{31}$$

The small number $\varsigma \in \left(0, \frac{1}{4}\right)$ in (29)-(30) is the same one in (26)-(28). As for λ_n, we may apply additional adjustment to $\lambda_{n,j}^{(1)}$ by replacing $\hat{\sigma}_\eta^{(1)}$ with $\hat{\sigma}_\eta^{(2)}$, which may result better performance within small samples.

In Lemmas B.5 and B.6, we show

$$\begin{aligned}
\hat{\sigma}_\eta^{(1)} - \sigma_\eta &= o_p(1), &(32)\\
\max_j \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[Z_{ij}\hat{\pi}_j^{(2)} - Z_{ij}\pi_j^*\right]^2} &= O_p\left(\left(\sqrt{\sigma_\eta^2 k_1} \vee \sqrt{\sigma_\eta}\right)\left(\frac{\log(p \vee d)}{n}\right)^{\frac{1}{2}-\varsigma}\right), &(33)\\
\hat{\sigma}_\epsilon^{(1)} - \sigma_\epsilon &= o_p(1), &(34)
\end{aligned}$$

provided that

$$\begin{aligned}
\sigma_\eta &= o\left(\left(\frac{n}{\log(p \vee d)}\right)^{\frac{1}{4}}\right), &(35)\\
\sigma_\epsilon &= o\left(\left(\frac{n}{\log(p \vee d)}\right)^{\frac{1}{4}}\right). &(36)
\end{aligned}$$

Consequently, for the estimates, $\hat{\beta}^{(2)}$, returned by Step 2 of the *modified* algorithm based on $\left|\hat{\beta}^{(1)}\right|_1$, Lemma B.6 gives

$$\begin{aligned}
\left|\hat{\beta}^{(2)} - \beta^*\right|_2 &= O_p\left(\bar{B}^{(2)}\right), &(37)\\
\left|\hat{\beta}^{(2)} - \beta^*\right|_1 &= O_p\left(\sqrt{k_2}\bar{B}^{(2)} + |\beta^*_{S_\tau^c}|_1\right), &(38)
\end{aligned}$$

where

$$\begin{aligned}
\bar{B}^{(2)} &:= \frac{\sqrt{k_2}}{\underline{\kappa}_2}\mathcal{T}_0^{(1)} + \sqrt{\frac{\mathcal{T}_0^{(1)}|\beta^*_{S_\tau^c}|_1}{\underline{\kappa}_2}},\\
\mathcal{T}_0^{(1)} &:= \max\left\{\left(|\beta^*|_1 \vee \sigma_\epsilon\right)\mathcal{T}_1^f,\ \sigma_\eta|\beta^*|_1\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\varsigma},\ \sigma_\epsilon\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\varsigma}\right\},\\
\mathcal{T}_1^f &:= \left(\sqrt{\sigma_\eta^2 k_1} \vee \sqrt{\sigma_\eta}\right)\left(\frac{\log(p \vee d)}{n}\right)^{\frac{1}{2}-\varsigma}.
\end{aligned}$$

Note that if σ_η, $\sigma_\epsilon = O(1)$, the right-hand-sides in (37) and (38) are bounded from above by the right-hand-sides in (24) and (25), respectively. Since the *modified* algorithm only requires (35) and (36) rather than σ_η, $\sigma_\epsilon = O(1)$ in Theorem 2.2, we expect it to work better within small samples when the noise variances are relatively high.

In the following section, we turn to Monte-Carlo simulation experiments and evaluate the small sample performance of our H2SLS where the second-stage regularization parameter is chosen according to the *modified* algorithm introduced above.

3 Simulations

We generate the data based on (1) and (2) where Z_i is a $p \times d$ matrix of independent standard normal random variables, and Z_{ij} is independent of $(\epsilon_i, \eta_{i1}, ..., \eta_{ip})$ for all $j = 1, ..., p$. We choose $d = 400$ and $p = 400$. A hundred sets of *i.i.d.* $(Y_i, X_i, Z_i, \epsilon_i, \eta_i)_{i=1}^n$ are simulated where n is the sample size in each set and

$$(\epsilon_i, \eta_i) \sim_{i.i.d.} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \varrho\sigma_\epsilon\sigma_\eta & \cdots & \cdots & \varrho\sigma_\epsilon\sigma_\eta \\ \varrho\sigma_\epsilon\sigma_\eta & \sigma_\eta^2 & 0 & \cdots & 0 \\ \vdots & 0 & \sigma_\eta^2 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \varrho\sigma_\epsilon\sigma_\eta & 0 & \cdots & 0 & \sigma_\eta^2 \end{pmatrix}\right) \tag{39}$$

with $\sigma_\epsilon^2 := \text{var}(\epsilon_i)$, $\sigma_\eta^2 := \text{var}(\eta_{ij})$ for all j, and ϱ the correlation between ϵ_i and η_{ij}. We set $\varrho = 0.05$ to introduce endogeneity in all 400 components of X_i while ensuring the covariance matrix in (39) generated by Matlab to be positive definite for the choices of σ_ϵ and σ_η in Table 3.1 (larger values of ϱ fail to maintain the positive definiteness of (39)).

Table 3.1: Parameters for Designs A, B, C

Parameters	**Exp. 1**	**Exp. 2**	**Exp. 3**	**Exp. 4**	**Exp. 5**
$\beta_j^* (j = 1, ..., 4)$	0.5	0.5	0.25	0.5	0.5
σ_ϵ	0.5	1	0.5	0.5	0.5
σ_η	0.5	1	0.5	0.5	0.5
n	399	399	399	200	800

Three sparse designs are considered. In terms of the first-stage equations' coefficients, for every j and $l = 5, ..., 400$, Design A sets $\pi_{jl}^* = 0$, Design B sets $\pi_{jl}^* = \frac{0.1}{l}$, and Design C sets $\pi_{jl}^* = 0.25^{l-3}$; for all three designs, $\pi_{jl}^* = 0.5$ for every j and $l = 1, ..., 4$. In terms of the main equation's coefficients, for $j = 5, ..., 400$, Design A sets $\beta_j^* = 0$, Design B sets $\beta_j^* = \frac{0.1}{j}$, and Design C sets $\beta_j^* = 0.25^{j-3}$. For each sparse design, we perform five experiments differing in β_j^* $(j = 1, ..., 4)$, σ_ϵ, σ_η, and n. Table 3.1 summarizes the parameters for each of the five experiments.

For each simulation run $h = 1, ..., 100$, we apply the *modified* algorithm in Section 2.2 with $\varsigma = \frac{1}{256}$. For $\lambda_{n,j}$s in (3), we apply (29)-(30) and iterate the "Adjusted-Penalization" step three times (i.e., a total of four iterations including the "Over-Penalization" step). With $\hat{\sigma}_\eta^{(4)}$ from the last iteration, we set $\hat{\sigma}_\eta := \hat{\sigma}_\eta^{(4)}$, which is used in the *modified* algorithm for selecting λ_n in (4). For λ_n, we apply (26)-(28) with $C = 0.5$ in (28) and iterate the "Adjusted-Penalization" step twice (i.e., a total of three iterations including the "Over-Penalization" step). Let λ_n^h denote the final second-stage regularization parameter and $\hat{\beta}^h$ the second-stage estimate for β^* in the hth run. Tables 3.2-3.4 display the mean of λ_n^hs, $\frac{1}{100}\sum_{h=1}^{100}\lambda_n^h$, the mean of the l_0-norms of $\hat{\beta}^h$, $\frac{1}{100}\sum_{h=1}^{100}\sum_{j=1}^{400} 1\left\{\hat{\beta}_j^h \neq 0\right\}$, the mean of the l_2-errors, $\frac{1}{100}\sum_{h=1}^{100}\left|\hat{\beta}^h - \beta^*\right|_2$, as well as the mean of the l_1-errors, $\frac{1}{100}\sum_{h=1}^{100}\left|\hat{\beta}^h - \beta^*\right|_1$, for Designs A, B, and C, respectively.

The results show that our H2SLS in conjunction with the *modified* algorithm for setting λ_n and $\lambda_{n,j}$s perform well for these sparse designs. The directions and magnitudes of the changes in the results from Experiment 1 to another experiment agree with our predictions based on (37) and (28).

For Design A (the exact sparsity case), the bound in (37) can be reduced to $O\left(\underline{\kappa}_2^{-1}\sqrt{k_2}\mathcal{T}_0^{(1)}\right)$, a term that accounts for the estimation error; consequently, in view of (28), when the noise variance, σ_η, is doubled, the means of the λ_n^hs and l_2-errors are approximately doubled; when $|\beta^*|_1$ is changed from 2 to 1, the means of the λ_n^hs and l_2-errors are also nearly halved; when the sample size n is nearly doubled (halved), the means of the λ_n^hs and l_2-errors are nearly decreased by a factor of $\sqrt{2}$ (respectively, increased by a factor of $\sqrt{2}$).

For the approximately sparse designs B and C, similar patterns are witnessed even though an additional term, $\sqrt{\underline{\kappa}_2^{-1}|\beta^*_{S_\tau^c}|_1\mathcal{T}_0^{(1)}}$, appears in (37) to account for the approximation error. We have computed that $\sum_{j=5}^{400}\left|\beta_j^*\right| \approx 0.449 \leq k_1\sqrt{\frac{\log(p\vee d)}{n}} = 0.490$ for Design B and $\sum_{j=5}^{400}\left|\beta_j^*\right| \approx 0.083 \leq 0.490$ for Design C; theoretically speaking, this would mean condition (18) in Corollary 2.1 and Lemma B.5 is satisfied. The fact that the l_2-errors of Design C are similar to those of Design B suggests that the actual approximation errors are likely to be much smaller than the actual estimation errors. On the other hand, Design B yields the highest mean of the l_1-errors, followed by Design C. In view of (38), this is because B has the largest $\sum_{j=5}^{400}\left|\beta_j^*\right|$ among all three designs.

Exp	**Table 3.2: Design A**			
#	**mean** λ_n	**mean** $\left\|\hat{\beta}\right\|_0$	**mean** $\left\|\hat{\beta}-\beta^*\right\|_2$	**mean** $\left\|\hat{\beta}-\beta^*\right\|_1$
1	0.173	4	0.264	0.512
2	0.280	4.570	0.446	0.851
3	0.084	6.240	0.135	0.268
4	0.231	4.050	0.369	0.701
5	0.132	4	0.196	0.380

Exp	**Table 3.3: Design B**			
#	**mean** λ_n	**mean** $\left\|\hat{\beta}\right\|_0$	**mean** $\left\|\hat{\beta}-\beta^*\right\|_2$	**mean** $\left\|\hat{\beta}-\beta^*\right\|_1$
1	0.176	4	0.273	0.969
2	0.290	4.420	0.470	1.342
3	0.084	6.740	0.144	0.720
4	0.232	4.080	0.373	1.152
5	0.134	4	0.205	0.836

Exp	**Table 3.4: Design C**			
#	**mean** λ_n	**mean** $\left\|\hat{\beta}\right\|_0$	**mean** $\left\|\hat{\beta}-\beta^*\right\|_2$	**mean** $\left\|\hat{\beta}-\beta^*\right\|_1$
1	0.176	4	0.278	0.606
2	0.284	4.480	0.460	0.951
3	0.085	6.620	0.150	0.352
4	0.232	4.070	0.376	0.787
5	0.136	4	0.214	0.478

4 Future directions

This paper has explored the validity of the H2SLS estimation for linear models where the number of endogenous regressors in the main equation and the number of instruments in the first-stage equations can exceed the sample size n, and the regression coefficients are sufficiently sparse. We establish finite-sample performance bounds and also provide a simple method for choosing the

regularization parameter with asymptotic guarantees. The proposed procedure is tested on simulated data and the results show that our H2SLS in conjunction with the method for setting the regularization parameters perform well for various sparse designs.

There are two extensions that worth exploring in the future research. First, as we have discussed in Section 2.2, selecting the constant C in (21) can be assisted with the CV criterion or the ESCV criterion proposed by Lim and Yu (2013). Here we lay out the details on how the ESCV criterion can be tailored to our "Adjusted-Penalization" step. Let the n observations be randomly assigned into T subsamples of size $(n-L)$, where $L = \lfloor \frac{n}{T} \rfloor$. Suppose we consider a set of C^ms $(m = 1, ..., M)$ for the constant C in (21) and denote the resulting λ_n as λ_n^m for each choice C^m. Given λ_n^m and the subsample t, the "Adjusted-Penalization" step is performed to obtain $\hat{\beta}_t(\lambda_n^m)$ and $\hat{Y}_t(\lambda_n^m) = \hat{X}\hat{\beta}_t(\lambda_n^m)$. For each $m = 1, ..., M$, following Lim and Yu (2013), we form

$$\mathrm{ES}(\lambda_n^m) := \frac{\widehat{\mathrm{Var}}(\hat{Y}(\lambda_n^m))}{\left|\bar{\hat{Y}}(\lambda_n^m)\right|_n^2} = \frac{L}{n-L}\frac{1}{\mathcal{Z}^2(\lambda_n^m)}$$

with

$$\begin{aligned}
\widehat{\mathrm{Var}}(\hat{Y}(\lambda_n^m)) &:= \frac{1}{T}\sum_{t=1}^{T}\left|\hat{Y}_t(\lambda_n^m) - \bar{\hat{Y}}(\lambda_n^m)\right|_n^2, \\
\mathcal{Z}^2(\lambda_n^m) &:= \frac{\bar{\hat{Y}}(\lambda_n^m)}{\sqrt{\frac{n-L}{L}\widehat{\mathrm{Var}}(\hat{Y}(\lambda_n^m))}}, \\
\bar{\hat{Y}}(\lambda_n^m) &:= \frac{1}{T}\sum_{t=1}^{T}\hat{Y}_t(\lambda_n^m),
\end{aligned}$$

where we denote $|a|_n^2 := \frac{1}{n}\sum_{i=1}^n a_i^2$. Let $\hat{\sigma}_{X_j^*} = \sqrt{\frac{1}{n}\sum_{i=1}^n \hat{X}_{ij}^2}$. We then apply their ESCV criterion: *Choose λ_n^m such that it minimizes* $\mathrm{ES}(\lambda_n^m)$ *over all m and $\sum_{j=1}^p \hat{\sigma}_{X_j^*}\left|\hat{\beta}_j(\lambda_n^m)\right|$ is no greater than the one resulting from the optimal Cross-Validation (CV) choice.* Lim and Yu (2013) recommend a grid-search algorithm to find a local minimum of ES as what is often done for the CV. Because the computational cost is rather high for our simulation exercise, we did not apply the ESCV criterion for selecting C in Section 4. However, it would be useful to evaluate the performance of this procedure with real data sets.

Second, it may be worthwhile to extend our analysis to allow non-sub-Gaussian errors ϵ and η in (1) and (2). There are a couple of ways to relax the sub-Gaussian condition on the error terms. For example, the square-root Lasso (as in Belloni, Chernozhukov, and Wang, 2014) and the pivotal Dantzig selector (as in Gautier and Tsybakov, 2014) allow these authors to evoke a bound for moderate deviations of self-normalized sums of random variables, which do not require sub-Gaussian tails. However, compared to the standard Lasso, the square-root Lasso or the pivotal Dantzig selector involves a more sophisticated optimization algorithm computation-wise. Another paper by Minsker (2014) that uses a "trick" originally noted in Nemirovski and Yudin (1983) is also able to avoid imposing a sub-Gaussian condition on the error terms. It is possible to apply these techniques in our problem, albeit doing so would distract the main focus of this paper; therefore, we leave these extensions to future research.

Besides the above extensions, we discuss two important future directions beyond this research. One direction regards the high dimensional "control function" approach, which is a close alternative to the H2SLS. Another direction regards inference strategies that can be built upon the H2SLS.

The "control function" approach. As an alternative to the $\hat{\beta}_{H2SLS}$ proposed in this paper, another type of two-stage estimator based on the "control function" approach is worth being explored. The "control function" approach includes the first-stage estimation residuals $\hat{\eta}_{ij} = X_{ij} - Z_{ij}\hat{\pi}_j$ as additional "control variables" (for the part of X_i that is correlated with ϵ_i) in the regression of Y_i on X_i. In particular, we can perform the following estimation

$$\hat{\beta}_{HCF} \in \text{argmin}_{\beta,\gamma\in\mathbb{R}^p} \frac{1}{2n}|Y - X\beta - \hat{\eta}\gamma|_2^2 + \lambda_n \left(|\beta|_1 + |\gamma|_1\right),$$

where the estimates $\hat{\eta} = (X_j - Z_j\hat{\pi}_j)_{j=1}^p$ of $\eta = \left(X_j - Z_j\pi_j^*\right)_{j=1}^p$ are obtained from (3).

When (1) and (2) are in the classical settings (fixed p and d), the two-stage least squares estimator is algebraically equivalent to a "control function" approach (e.g., Garen, 1984). Such algebraic equivalence no longer holds when regularization is introduced in the estimation. Nevertheless, the connection between $\hat{\beta}_{H2SLS}$ and $\hat{\beta}_{HCF}$ remains an interesting question for future research.

Inference based on H2SLS. Among existing literature, establishing variable selection consistency is the most popular approach to obtain inference results because it allows one to apply procedures from the classical low-dimensional regime by considering only the selected regressors. Variable selection consistency has been proved under an "incoherence" condition on the design matrix for the Lasso (e.g., Wainwright, 2009; Ravikumar, et al., 2010). The "incoherence condition" is a refined version of the "irrepresentable condition" by Zhao and Yu (2006) and the "neighborhood stability condition" by Meinshausen and Bühlmann (2006). Zhu (2013) establishes results regarding variable selection of $\hat{\beta}_{H2SLS}$, which could be of independent interest[1].

The drawback to the aforementioned post-variable-selection inference strategy is that the resulting estimators suffer the problems arising from the nonuniformity of limit theory (see, e.g., Leeb and Pötscher, 2006). Here we mean the nonuniformity in β^*, the parameter vector of interest. Among recent development, several uniform inference strategies have been proposed (e.g., Javanmard and Montanari, 2014; van de Geer, Bühlmann, Ritov, and Dezeure, 2014; Zhang and Zhang, 2014). For the models of our interest, these inference strategies can be applied to construct confidence intervals for any coefficient in (1). In particular, these strategies rely on an initial estimator and in our case, such a candidate can be the $\hat{\beta}^{(2)}$ in Theorem 2.2. To illustrate, we only sketch the strategy by Zhang and Zhang (2014) based on $\hat{\beta}^{(2)}$ in the following.

Denote X_{-j} the columns of X excluding the jth column. Following Zhang and Zhang (2014), for $j \in \{1, ..., p\}$, we construct the following "de-biased" estimator,

$$\tilde{\beta}_j := \hat{\beta}_j^{(2)} + \frac{r_j^T\left(Y - X\hat{\beta}^{(2)}\right)}{r_j^T X_j} \tag{40}$$

where $r_j = \hat{X}_j - \hat{X}_{-j}\hat{\theta}_j$ with

$$\hat{\theta}_j \in \arg\min_{\theta_j\in\mathbb{R}^{p-1}} \left\{\frac{|\hat{X}_j - \hat{X}_{-j}\theta_j|_2^2}{2n} + \mu_{n,j}\left|\theta_j\right|_1\right\},$$

for a non-negative tuning parameter $\mu_{n,j}$ of order $\sqrt{\frac{\log p}{n}}$. Note that (40) yields

$$\sqrt{n}\left(\tilde{\beta}_j - \beta_j^*\right) = \frac{\frac{1}{\sqrt{n}}r_j^T\epsilon}{\frac{1}{n}r_j^T X_j} - \frac{\frac{1}{\sqrt{n}}\sum_{l\neq j} r_j^T X_l\left(\hat{\beta}_l^{(2)} - \beta_l^*\right)}{\frac{1}{n}r_j^T X_j}. \tag{41}$$

[1]Note that in Zhu (2013), while the result establishes $J(\hat{\beta}_{H2SLS}) = J(\beta^*)$ with high probability for exactly sparse β^*, the argument follows through if $J(\beta^*)$ is replaced with the thresholded subset S_τ when β^* is approximately sparse.

Moreover, we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{l\neq j} r_j^T X_l\left(\hat{\beta}_l^{(2)}-\beta_l^*\right)\\ \leq\ & \max_{l\neq j}\frac{1}{\sqrt{n}}\left[\left|r_j^T\hat{X}_l\right|+\left|r_j^T\left(X_l-\hat{X}_l\right)\right|\right]\left(\left|\hat{\beta}^{(2)}-\beta^*\right|_1\right).\end{aligned}$$

We can apply the argument in Zhang and Zhang (2014, Proposition 1) to show that

$$\max_{l\neq j}\frac{1}{n}\left|r_j^T\hat{X}_l\right| = O_p\left(\sqrt{\frac{\log p}{n}}\right).$$

By Lemma B.7 in this paper, we also have

$$\frac{1}{n}\max_{l\neq j}\left|r_j^T\left(X_l-\hat{X}_l\right)\right| = O_p\left(\mathcal{E}\right)$$

where

$$\mathcal{E} := \left(\left|\hat{\theta}_j\right|_1\vee 1\right)\max\left\{\sigma_{X^*}\mathcal{T}_1,\ \rho_{X^*}\rho_\eta\sqrt{\frac{\log p}{n}}\right\}.$$

Note that, under the conditions in Theorem 2.2, if $\underline{\kappa}_2^{-1}=O(1)$ and

$$|\beta^*_{S_\tau^c}|_1 = O\left(\left(|\beta^*|_1\vee 1\right)\left(k_2\vee 1\right)\sqrt{\frac{(k_1\vee 1)\log(d\vee p)}{n}}\right),$$

then we have

$$\left|\hat{\beta}^{(2)}-\beta^*\right|_1 = O_p\left(\left(|\beta^*|_1\vee 1\right)\left(k_2\vee 1\right)\left(\frac{(k_1\vee 1)\log(d\vee p)}{n}\right)^{\frac{1}{2}-\varsigma}\right).$$

Putting these facts together, if

$$\sqrt{n}\left(\mathcal{E}\vee\sqrt{\frac{\log p}{n}}\right)\left(|\beta^*|_1\vee 1\right)\left(k_2\vee 1\right)\left(\frac{(k_1\vee 1)\log(d\vee p)}{n}\right)^{\frac{1}{2}-\varsigma} = o(1),$$

then

$$\frac{1}{\sqrt{n}}\sum_{l\neq j} r_j^T X_l\left(\hat{\beta}_l^{(2)}-\beta_l^*\right) = o_p(1).$$

Consequently, if $\frac{1}{n}r_j^T X_j \xrightarrow{p} D\neq 0$, then $\sqrt{n}\left(\tilde{\beta}_j-\hat{\beta}_j^{(2)}\right)$ has the same asymptotic distribution as the leading term $D^{-1}\frac{r_j^T\epsilon}{\sqrt{n}}$ in (41).

Note that the de-biased estimator $\tilde{\beta}_j$ in (40) relies on $\hat{\beta}_j^{(2)}$ whose construction uses $\hat{k}_1 = \max_{j=1,\ldots,p}|J(\hat{\pi}_j)|$. To ensure $\hat{k}_1\geq k_1$ with probability at least $1-o(1)$, we impose a condition on $\min_{l\in S_{\tau_j}}|\pi^*_{jl}|$ in Lemma B.4. Under such a condition, the de-biased estimator discussed above is valid uniformly in β^* only but not in the nuisance parameters, π_j^*s. Developing a de-biased H2SLS procedure that is valid uniformly in both β^* and π_j^*s would be worth exploring in the future research.

A Appendix: Main proofs

A.1 Proof for Theorem 2.1

Lemma A.1. Suppose λ_n satisfies that $\lambda_n \geq \mathcal{T}_0$ and the conditions in Lemmas A.3-A.4 hold. Let $b_0 = \underline{\kappa}_2 \left(\frac{\tilde{\rho}^4_{X^*}}{\underline{\kappa}_2^2} \vee 1 \right)$. If

$$|\beta^*|_1 \lambda_n^{-1} \left(\frac{b_0 \log p}{n} \vee \mathcal{T}_1^2 \right) \leq c'' \tag{42}$$

for some universal constant $c'' > 0$, then there exist positive universal constants c_0^*, c_1^* and c_2^* such that, for $\tau = \frac{\lambda_n}{\underline{\kappa}_2}$ in (12), we have

$$\begin{aligned} |\hat{\beta}_{H2SLS} - \beta^*|_2 &\leq c_0^* \left(\frac{\lambda_n \sqrt{k_2}}{\underline{\kappa}_2} + \sqrt{\frac{\lambda_n |\beta^*_{S_\tau^c}|_1}{\underline{\kappa}_2}} \right) := \bar{B}, \\ |\hat{\beta}_{H2SLS} - \beta^*|_1 &\leq 4 \left(\sqrt{k_2} \bar{B} + |\beta^*_{S_\tau^c}|_1 \right), \end{aligned}$$

with probability at least $1 - c_1^* \exp(-c_2^* \log p)$.

Proof. We write

$$\begin{aligned} Y &= X\beta^* + \epsilon = X^*\beta^* + (X\beta^* - X^*\beta^* + \epsilon) \\ &= X^*\beta^* + (\eta\beta^* + \epsilon) \\ &= \hat{X}\beta^* + (X^* - \hat{X})\beta^* + \eta\beta^* + \epsilon \\ &= \hat{X}\beta^* + \xi, \end{aligned}$$

where

$$\xi := (X^* - \hat{X})\beta^* + \eta\beta^* + \epsilon.$$

Let $\hat{v}^0 = \hat{\beta}_{H2SLS} - \beta^*$. Given a set S, recall that $\hat{v}_S \in p \times 1$ is the vector that has the same coordinates as $\hat{v}$ on S and zero coordinates on the complement S^c of S. Define the Lagrangian $L(\beta;\, \lambda_n) = \frac{1}{2n}|Y - \hat{X}\beta|_2^2 + \lambda_n |\beta|_1$. Since $\hat{\beta}_{H2SLS}$ is optimal, we have

$$L(\hat{\beta}_{H2SLS};\, \lambda_n) \leq L(\beta^*;\, \lambda_n) = \frac{1}{2n}|\xi|_2^2 + \lambda_n |\beta^*|_1,$$

which yields

$$\begin{aligned} 0 \leq \frac{1}{2n}|\hat{X}\hat{v}^0|_2^2 &\leq \frac{1}{n}\xi^T \hat{X}\hat{v}^0 + \lambda_n \left\{ |\beta^*_{S_\tau}|_1 + |\beta^*_{S_\tau^c}|_1 - |(\beta^*_{S_\tau} + \hat{v}^0_{S_\tau},\, \beta^*_{S_\tau^c} + \hat{v}^0_{S_\tau^c})|_1 \right\} \\ &\leq |\hat{v}^0|_1 |\frac{1}{n}\hat{X}^T\xi|_\infty + \lambda_n \left\{ |\hat{v}^0_{S_\tau}|_1 - |\hat{v}^0_{S_\tau^c}|_1 + 2|\beta^*_{S_\tau^c}|_1 \right\} \qquad (43) \\ &\leq |\hat{v}^0|_1 |\frac{1}{n}\hat{X}^T\xi|_\infty + \frac{\lambda_n}{2} \left\{ 2|\hat{v}^0_{S_\tau}|_1 - 2|\hat{v}^0_{S_\tau^c}|_1 + 4|\beta^*_{S_\tau^c}|_1 \right\} \\ &\leq \frac{\lambda_n}{2} \left\{ 3|\hat{v}^0_{S_\tau}|_1 - |\hat{v}^0_{S_\tau^c}|_1 + 4|\beta^*_{S_\tau^c}|_1 \right\}, \qquad (44) \end{aligned}$$

where (44) holds since $|\hat{v}^0|_1 = |\hat{v}^0_{S_\tau}|_1 + |\hat{v}^0_{S_\tau^c}|_1$, $\lambda_n \geq \mathcal{T}_0$ and $\mathcal{T}_0 \geq 2 \left| \frac{\hat{X}^T\xi}{n} \right|_\infty$ (by Lemma A.4) with probability at least $1 - c_7 \exp(-c_8 \log p)$; consequently,

$$|\hat{v}^0|_1 \leq 4|\hat{v}^0_{S_\tau}|_1 + 4|\beta^*_{S_\tau^c}|_1 \leq 4\sqrt{k_2}|\hat{v}^0|_2 + 4|\beta^*_{S_\tau^c}|_1. \tag{45}$$

We bound the cardinality of S_τ from above in terms of the threshold $\tau = \frac{\lambda_n}{\underline{\kappa}_2}$. Note that we have

$$\sum_{j=1}^{p} \left|\beta_j^*\right| \geq \sum_{j \in S_\tau} \left|\beta_j^*\right| \geq \tau k_2$$

and therefore $k_2 \leq \tau^{-1}|\beta^*|_1$. Putting the pieces together yields

$$|\hat{v}^0|_1 \leq 4\sqrt{\tau^{-1}|\beta^*|_1}|\hat{v}^0|_2 + 4|\beta^*_{S_\tau^c}|_1.$$

By the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$|\hat{v}^0|_1^2 \leq 32\tau^{-1}|\beta^*|_1|\hat{v}^0|_2^2 + 32|\beta^*_{S_\tau^c}|_1^2. \tag{46}$$

By substituting (46) into condition (54) from Lemma A.3, we obtain

$$\left|\hat{v}^{0T}\frac{\hat{X}^T\hat{X}}{n}\hat{v}^0\right| \geq |\hat{v}^0|_2^2\left\{\frac{\underline{\kappa}_2}{4} - c^{'}|\beta^*|_1\tau^{-1}\left(\mathcal{T}_1^2 \vee \frac{b_0 \log p}{n}\right)\right\} - c^{'}|\beta^*_{S_\tau^c}|_1^2\left(\mathcal{T}_1^2 \vee \frac{b_0 \log p}{n}\right), \tag{47}$$

for some positive universal constant $c^{'}$, with probability at least

$$1 - c_5\exp(-c_6 \log p) - c_7\exp(-c_8 \log p) \geq 1 - c_1^*\exp(-c_2^* \log p)$$

where $b_0 = \underline{\kappa}_2\left(\frac{\tilde{\rho}^4_{X^*}}{\underline{\kappa}_2^2} \vee 1\right)$, $c_1^* = c_5 + c_7$ and $c_2^* = (c_6 \wedge c_8)$.

We now proceed case by case. Let

$$\delta^* := 4c^{'\frac{1}{2}}\underline{\kappa}_2^{-\frac{1}{2}}|\beta^*_{S_\tau^c}|_1\sqrt{\mathcal{T}_1^2 \vee \frac{b_0 \log p}{n}}.$$

Provided that $16c^{'}|\beta^*_{S_\tau^c}|_1\left(\mathcal{T}_1^2 \vee \frac{b_0 \log p}{n}\right) \leq \lambda_n$ (which is guaranteed by condition (42)), we have

$$\delta^* \leq \sqrt{\frac{\lambda_n|\beta^*_{S_\tau^c}|_1}{\underline{\kappa}_2}} \leq \frac{\lambda_n\sqrt{k_2}}{\underline{\kappa}_2} + \sqrt{\frac{\lambda_n|\beta^*_{S_\tau^c}|_1}{\underline{\kappa}_2}}.$$

Case (i): If $\left|\hat{v}^0\right|_2 < \frac{\lambda_n\sqrt{k_2}}{\underline{\kappa}_2} + \sqrt{\frac{\lambda_n|\beta^*_{S_\tau^c}|_1}{\underline{\kappa}_2}}$, then we are done.

Case (ii): If $\left|\hat{v}^0\right|_2 \geq \frac{\lambda_n\sqrt{k_2}}{\underline{\kappa}_2} + \sqrt{\frac{\lambda_n|\beta^*_{S_\tau^c}|_1}{\underline{\kappa}_2}} \geq \delta^*$ so that $\frac{|v^0|_2^2}{16}\underline{\kappa}_2 \geq c^{'}|\beta^*_{S_\tau^c}|_1^2\left(\mathcal{T}_1^2 \vee \frac{b_0 \log p}{n}\right)$ in (47), under the condition $c^{'}|\beta^*|_1\tau^{-1}\left(\frac{b_0 \log p}{n} \vee \mathcal{T}_1^2\right) \leq \frac{\underline{\kappa}_2}{16}$ (which is guaranteed by condition (42) given $\tau = \frac{\lambda_n}{\underline{\kappa}_2}$), (47) implies that, with probability at least $1 - c_1^*\exp(-c_2^* \log p)$,

$$\left|\hat{v}^{0T}\frac{\hat{X}^T\hat{X}}{n}\hat{v}^0\right| \geq \underline{\kappa}_2\left\{\frac{|\hat{v}^0|_2^2}{4} - \frac{|\hat{v}^0|_2^2}{16} - \frac{|\hat{v}^0|_2^2}{16}\right\} = \frac{1}{8}\underline{\kappa}_2|\hat{v}^0|_2^2 > 0 \quad \text{for } \left|\hat{v}^0\right|_2 \neq 0, \tag{48}$$

which shows that $\left|\hat{v}^{0T}\frac{\hat{X}^T\hat{X}}{n}\hat{v}^0\right|$ is bounded away from zero. Now, from (43), since $\lambda_n \geq \mathcal{T}_0$ and $\mathcal{T}_0 \geq 2\left|\frac{\hat{X}^T\xi}{n}\right|_\infty$ (by Lemma A.4) with probability at least $1 - c_7\exp(-c_8 \log p)$,

$$\begin{aligned}
\frac{1}{2n}|\hat{X}\hat{v}^0|_2^2 &\leq |\hat{v}^0|_1|\frac{1}{n}\hat{X}^T\xi|_\infty + \lambda_n\left\{|\hat{v}^0_{S_\tau}|_1 - |\hat{v}^0_{S_\tau^c}|_1 + 2|\beta^*_{S_\tau^c}|_1\right\} \\
&\leq \lambda_n\left\{2|\hat{v}^0|_1 + 2|\beta^*_{S_\tau^c}|_1\right\} \\
&\leq c_0^{''}\lambda_n \max\left\{\sqrt{k_2}|\hat{v}^0|_2,\, |\beta^*_{S_\tau^c}|_1\right\}
\end{aligned}$$

for some positive universal constant c_0'', where we have used (45) in the last inequality.

If $\max\left\{\sqrt{k_2}|\hat{v}^0|_2,\, |\beta^*_{S^c_\tau}|_1\right\} = \sqrt{k_2}|\hat{v}^0|_2$, then by (48), we have

$$\frac{1}{16}\underline{\kappa}_2|v^0|_2^2 \leq \frac{1}{2n}|\hat{X}\hat{v}^0|_2^2 \leq c_0''\lambda_n\sqrt{k_2}|\hat{v}^0|_2$$

which implies

$$|\hat{v}^0|_2 \leq 16c_0''\frac{\lambda_n\sqrt{k_2}}{\underline{\kappa}_2}. \tag{49}$$

If $\max\left\{\sqrt{k_2}|\hat{v}^0|_2,\, |\beta^*_{S^c_\tau}|_1\right\} = |\beta^*_{S^c_\tau}|_1$, then

$$\frac{1}{16}\underline{\kappa}_2|v^0|_2^2 \leq \frac{1}{2n}|\hat{X}\hat{v}^0|_2^2 \leq c_0''\lambda_n|\beta^*_{S^c_\tau}|_1$$

which implies

$$|v^0|_2 \leq \sqrt{16c_0''}\sqrt{\frac{\lambda_n|\beta^*_{S^c_\tau}|_1}{\underline{\kappa}_2}}. \tag{50}$$

In view of case (i) along with (49)-(50), we have

$$\left|\hat{\beta}_{H2SLS} - \beta^*\right|_2 \leq c_0^*\left(\frac{\lambda_n\sqrt{k_2}}{\underline{\kappa}_2} + \sqrt{\frac{\lambda_n|\beta^*_{S^c_\tau}|_1}{\underline{\kappa}_2}}\right)$$

for some positive universal constant c_0^*. The bound on $\left|\hat{\beta}_{H2SLS} - \beta^*\right|_1$ then follows from (45). □

Lemma A.2. Suppose Assumption 2.5(i) holds. (a) Let $(Z_{ijl})_{i=1}^n$ be independent variables such that $\sup_{r\geq 1} r^{-\frac{1}{2}}\left(\mathbb{E}\left|Z_{ijl}\right|^r\right)^{\frac{1}{r}} \leq 1$ for all $i = 1, ..., n$, $j = 1, ..., p$ and $l = 1, ..., d$. Then,

$$\mathbb{P}\left(\max_{j,l}\left|\hat{\sigma}^2_{Z_{jl}} - \sigma^2_{Z_{jl}}\right| \leq 8e\sqrt{\frac{\log(pd)}{n}}\right) \geq 1 - 2\exp(-\log(pd)),$$

where $\hat{\sigma}^2_{Z_{jl}} = \frac{1}{n}\sum_{i=1}^n Z^2_{ijl}$ and $\sigma^2_{Z_{jl}} := \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n Z^2_{ijl}\right)$. (b) Moreover, under Assumptions 2.4 and 2.5(ii), there exist positive universal constants c_3, c_4 such that

$$\mathbb{P}\left(\max_{j=1,...,p}\left|\hat{\sigma}^2_{X^*_j} - \sigma^2_{X^*_j}\right| \leq 4\max_j \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X^{*2}_{ij}\right)\mathcal{T}_1\right) \geq 1 - c_3\exp(-c_4\log p),$$

where $\hat{\sigma}^2_{X^*_j} = \frac{1}{n}\sum_{i=1}^n \hat{X}^2_{ij}$ and $\sigma^2_{X^*_j} = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X^{*2}_{ij}\right)$.

Remark. Note that the bounds in Lemma A.2 imply (5) and (6).

Proof. For part (a), we apply Lemma B.1 with the choice $\varepsilon = 8e\sqrt{\frac{\log(pd)}{n}}$ and a union bound to obtain

$$\max_{j,l}\frac{1}{n}\sum_{i=1}^n Z^2_{ijl} \leq \max_{j,l}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n Z^2_{ijl}\right) + 8e\sqrt{\frac{\log(pd)}{n}}$$

with probability at least $1 - 2pd\exp(-2\log(pd)) = 1 - 2\exp(-\log(pd))$, where we have used the fact that $\sqrt{\frac{\log(pd)}{n}} \leq \frac{1}{2}$ implied by the first item in Assumption 2.5(i) (so that the term E_1 is no greater than the term E_2 in bound (57)).

For part (b), we provide a proof for a more general result, which is useful for proving Lemma A.3 later on. Note that we have

$$\begin{aligned}\left|\frac{\hat{X}^T\hat{X}-X^{*T}X^*}{n}\right|_\infty &\leq \left|\frac{X^{*T}(\hat{X}-X^*)}{n}\right|_\infty + \left|\frac{(\hat{X}-X^*)^T\hat{X}}{n}\right|_\infty \\ &\leq \left|\frac{X^{*T}(\hat{X}-X^*)}{n}\right|_\infty + \left|\frac{(\hat{X}-X^*)^TX^*}{n}\right|_\infty + \left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty \quad (51)\end{aligned}$$

To bound the term $\left|\frac{X^{*T}(\hat{X}-X^*)}{n}\right|_\infty$, first note that by Assumption 2.4, we have

$$\max_{j=1,\ldots,p}\sqrt{\frac{1}{n}\sum_{i=1}^n\left[Z_{ij}(\hat{\pi}_j-\pi_j^*)\right]^2} \leq \mathcal{T}_1$$

with probability at least $1-c_1\exp(-c_2\log(d\vee p))$; applying Lemma B.1 with the choice $\varepsilon = 8e\rho^2_{X^*}\sqrt{\frac{\log p}{n}}$ and a union bound, we have

$$\begin{aligned}\max_{j'}\frac{1}{n}\sum_{i=1}^n X_{ij'}^{*2} &\overset{(1)}{\leq} \max_{j'}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_{ij'}^{*2}\right) + 8e\rho^2_{X^*}\sqrt{\frac{\log p}{n}} \\ &\overset{(2)}{\leq} 2\max_{j'}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_{ij'}^{*2}\right)\end{aligned}$$

with probability at least $1-2p\exp(-2\log p) = 1-2\exp(-\log p)$, where by Assumption 2.5(i), (1) follows from the condition $\sqrt{\frac{\log p}{n}}\leq\frac{1}{2}$ (so that the term E_1 is no greater than the term E_2 in bound (57)) and (2) follows from the condition $8e\rho^2_{X^*}\sqrt{\frac{\log p}{n}} \leq \max_{j'}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_{ij'}^{*2}\right)$. As a consequence, we apply a Cauchy-Schwarz inequality and obtain, with probability at least $1-2\exp(-\log p) - c_1\exp(-c_2\log(p\vee d)) \geq 1-c_3\exp(-c_4\log p)$,

$$\begin{aligned}\max_{j',j}\left|\frac{1}{n}X_{j'}^{*T}(\hat{X}_j-X_j^*)\right| &= \max_{j',j}\left|\frac{1}{n}\sum_{i=1}^n X_{ij'}^*Z_{ij}(\hat{\pi}_j-\pi_j^*)\right| \\ &\leq \max_{j',j}\sqrt{\frac{1}{n}\sum_{i=1}^n X_{ij'}^{*2}}\sqrt{\frac{1}{n}\sum_{i=1}^n\left[Z_{ij}(\hat{\pi}_j-\pi_j^*)\right]^2} \\ &\leq \max_{j'}\sqrt{2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_{ij'}^{*2}\right)}\mathcal{T}_1, \quad (52)\end{aligned}$$

which bounds the term $\left|\frac{X^{*T}(\hat{X}-X^*)}{n}\right|_\infty$. To bound the term $\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty$, we again apply a Cauchy-Schwarz inequality and obtain

$$\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty \leq \mathcal{T}_1^2 \quad (53)$$

with probability at least $1-c_1\exp(-c_2\log(p\vee d))$.

Putting everything together, if $\mathcal{T}_1 \leq \max_j\sqrt{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_{ij}^{*2}\right)}$ (which is implied by Assumption 2.5(ii)), we have

$$\left|\frac{\hat{X}^T\hat{X}-X^{*T}X^*}{n}\right|_\infty \leq 4\max_j\sqrt{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_{ij}^{*2}\right)}\mathcal{T}_1$$

with probability at least $1-c_3\exp(-c_4\log p)$. The bound above implies the second claim in Lemma A.2. $\square$

Lemma A.3 (LRE condition). Under Assumption 2.3 and the conditions in Lemma A.2, there exist positive universal constants c^*, c_5, c_6 such that

$$\frac{|\hat{X}v^0|_2^2}{n} \geq \frac{\underline{\kappa}_2}{4}\left|v^0\right|_2^2 - \frac{c^*}{2}\underline{\kappa}_2\left(\frac{\tilde{\rho}^4_{X^*}}{\underline{\kappa}_2^2}\vee 1\right)\frac{\log p}{n}|v^0|_1^2 - \mathcal{T}_1^2\left|v^0\right|_1^2, \tag{54}$$

for any $v^0\in\mathbb{R}^p$, with probability at least $1-c_5\exp(-c_6\log p)$.

Proof. Note that by the elementary inequality $(a+b)^2\leq 2a^2+2b^2$, we have

$$\begin{aligned}\left|v^{0T}\frac{\hat{X}^T\hat{X}}{n}v^0\right| &\geq \left|v^{0T}\frac{X^{*T}X^*}{2n}v^0\right| - \left|v^{0T}\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}v^0\right| \\ &\geq \left|v^{0T}\frac{X^{*T}X^*}{2n}v^0\right| - \left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty\left|v^0\right|_1^2.\end{aligned}$$

We apply (53) and Assumption 2.3 to obtain

$$\begin{aligned}\left|v^{0T}\frac{\hat{X}^T\hat{X}}{n}v^0\right| &\geq \left|v^{0T}\frac{X^{*T}X^*}{2n}v^0\right| - \mathcal{T}_1^2\left|v^0\right|_1^2 \\ &\geq \frac{\underline{\kappa}_2}{4}\left|v^0\right|_2^2 - \frac{c^*}{2}\underline{\kappa}_2\left(\frac{\tilde{\rho}^4_{X^*}}{\underline{\kappa}_2^2}\vee 1\right)\frac{\log p}{n}|v^0|_1^2 - \mathcal{T}_1^2\left|v^0\right|_1^2\end{aligned}$$

with probability at least $1-c_1\exp(-c_2\log(p\vee d)) - 2\exp(-\log p)\geq 1-c_5\exp(-c_6\log p)$. $\square$

Lemma A.4 (Upper bound on $|\frac{1}{n}\hat{X}^T\xi|_\infty$). Under Assumptions 2.1, 2.2, 2.4, and 2.5, there exist positive universal constants c_7 and c_8 such that $2\left|\frac{\hat{X}^T\xi}{n}\right|_\infty \leq \mathcal{T}_0$ with probability at least $1-c_7\exp(-c_8\log p)$.

Proof. Recall that we have

$$\begin{aligned}\frac{1}{n}\hat{X}^T\xi &= \frac{1}{n}\hat{X}^T\left[(X^*-\hat{X})\beta^* + \eta\beta^* + \epsilon\right] \\ &= \frac{1}{n}\hat{X}^T(X^*-\hat{X})\beta^* + \frac{1}{n}X^{*T}\left[\eta\beta^*+\epsilon\right] + \frac{1}{n}(\hat{X}-X^*)^T\left[\eta\beta^*+\epsilon\right].\end{aligned}$$

Hence,

$$\begin{aligned}|\frac{1}{n}\hat{X}^T\xi|_\infty &\leq |\frac{1}{n}\hat{X}^T(\hat{X}-X^*)\beta^*|_\infty + |\frac{1}{n}X^{*T}\eta\beta^*|_\infty + |\frac{1}{n}X^{*T}\epsilon|_\infty \\ &\quad + |\frac{1}{n}(\hat{X}-X^*)^T\eta\beta^*|_\infty + |\frac{1}{n}(\hat{X}-X^*)^T\epsilon|_\infty.\end{aligned}$$

We need to bound each of the terms on the right-hand-side of the above inequality. Let us first bound $|\frac{1}{n}\hat{X}^T(\hat{X}-X^*)\beta^*|_\infty$. We have

$$\frac{1}{n}\hat{X}^T(\hat{X}-X^*)\beta^* = \begin{bmatrix}\sum_{j=1}^p\beta_j^*\frac{1}{n}\sum_{i=1}^n\hat{X}_{i1}(\hat{X}_{ij}-X^*_{ij}) \\ \vdots \\ \sum_{j=1}^p\beta_j^*\frac{1}{n}\sum_{i=1}^n\hat{X}_{ip}(\hat{X}_{ij}-X^*_{ij})\end{bmatrix}.$$

For any $j^{'} = 1, ..., p$, we have

$$\begin{aligned} |\sum_{j=1}^{p} \beta_j^* \frac{1}{n} \sum_{i=1}^{n} \hat{X}_{ij'}(\hat{X}_{ij} - X_{ij}^*)| &\leq \max_{j^{'}, j} |\frac{1}{n} \sum_{i=1}^{n} \hat{X}_{ij'}(\hat{X}_{ij} - X_{ij}^*)||\beta^*|_1 \\ &= \left| \frac{\hat{X}^T(\hat{X} - X^*)}{n} \right|_\infty |\beta^*|_1. \end{aligned}$$

Note that by (52) and (53),

$$\begin{aligned} \left| \frac{\hat{X}^T(\hat{X} - X^*)}{n} \right|_\infty &\leq \left| \frac{X^{*T}(\hat{X} - X^*)}{n} \right|_\infty + \left| \frac{(\hat{X} - X^*)^T(\hat{X} - X^*)}{n} \right|_\infty \\ &\leq 3 \max_j \sqrt{\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} X_{ij}^{*2} \right)} \mathcal{T}_1. \end{aligned}$$

Consequently,

$$|\frac{1}{n} \hat{X}^T(\hat{X} - X^*)\beta^*|_\infty \leq 3 \max_j \sqrt{\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} X_{ij}^{*2} \right)} \mathcal{T}_1 |\beta^*|_1,$$

with probability at least $1 - c_3 \exp(-c_4 \log p)$.

Applying Lemma B.1 with the choice $\varepsilon = 8e\rho_\epsilon^2 \sqrt{\frac{\log p}{n}}$, we have

$$\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \overset{(1)}{\leq} \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \right) + 8e\rho_\epsilon^2 \sqrt{\frac{\log p}{n}} \overset{(2)}{\leq} 2\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \right)$$

with probability at least $1 - 2p \exp(-2 \log p)$, where by Assumption 2.5(i), (1) follows from the condition $\sqrt{\frac{\log p}{n}} \leq \frac{1}{2}$ (so that the term E_1 is no greater than the term E_2 in bound (57)) and (2) follows from the condition $8e\rho_\epsilon^2 \sqrt{\frac{\log p}{n}} \leq \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \right)$. For the term $|\frac{1}{n}(X^* - \hat{X})^T \epsilon|_\infty$, we apply similar argument used for bounding $\left| \frac{1}{n} X_{j'}^{*T}(\hat{X}_j - X_j^*) \right|_\infty$ and obtain

$$\begin{aligned} |\frac{1}{n}(X^* - \hat{X})^T \epsilon|_\infty &\leq \max_j \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[Z_{ij}(\hat{\pi}_j - \pi_j^*) \right]^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2} \\ &\leq \mathcal{T}_1 \sqrt{2\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \right)} \end{aligned}$$

with probability at least $1 - 2\exp(-2\log p) - c_1 \exp(-c_2 \log(p \vee d)) \geq 1 - c_3 \exp(-c_4 \log p)$.

For the term $|\frac{1}{n} X^{*T} \eta \beta^*|_\infty$, we apply Lemma B.1 with the choice $\varepsilon = 4\sqrt{6} e \rho_{X^*} \rho_\eta \sqrt{\frac{\log p^2}{n}}$ and a union bound. Since $\mathbb{E}(\frac{1}{n} Z_{j'}^T \eta_j) = \mathbf{0}$ for all $j^{'}, j$, and the condition $\sqrt{\frac{\log p^2}{n}} \leq \frac{2}{5}$ implied by Assumption 2.5(i) (so that the term E_1 is no greater than the term E_2 in bound (57)), we have

$$\begin{aligned} |\frac{1}{n} X^{*T} \eta \beta^*|_\infty &\leq \max_{j^{'}, j} |\frac{1}{n} \sum_{i=1}^{n} X_{ij'}^* \eta_{ij}||\beta^*|_1 \qquad (55) \\ &\leq 8\sqrt{3} e \rho_{X^*} \rho_\eta \sqrt{\frac{\log p}{n}} |\beta^*|_1 \end{aligned}$$

with probability at least $1 - 2p^2 \exp(-3 \log p) = 1 - 2\exp(-\log p)$.

By Assumption 2.4, we have

$$\max_{j'} |\hat{\pi}_{j'} - \pi^*_{j'}|_1 \leq c' \left(\sqrt{k_1} err_e + \sqrt{k_1} err_a + \max_{j=1,\ldots,p} |\pi^*_{j,S^c_{\tau_j}}|_1 \right)$$

with probability at least $1 - c_1 \exp(-c_2 \log(p \vee d))$. Applying Lemma B.1 with $\varepsilon = 8e\sqrt{\frac{\rho_\eta^2 \log(dp^2)}{n}}$ and a union bound yields

$$\max_{j',j} |\frac{1}{n} \sum_{i=1}^n Z^T_{ij'} \eta_{ij}|_\infty \leq 8e \sqrt{\frac{\rho_\eta^2 \log(dp^2)}{n}}$$

with probability at least $1 - 2dp^2 \exp(-2\log(dp^2)) = 1 - 2\exp(-\log(dp^2))$, where we have used the fact that $\mathbb{E}(\frac{1}{n} Z^T_{j'} \eta_j) = \mathbf{0}$ for all j', j, and the condition $\sqrt{\frac{\log(dp^2)}{n}} \leq \frac{1}{2}$ implied by Assumption 2.5(i) (so that the term E_1 is no greater than the term E_2 in bound (57)). As a result, we have

$$\begin{aligned} & |\frac{1}{n}(X^* - \hat{X})^T \eta \beta^*|_\infty \qquad (56) \\ \leq \; & \max_{j'} |\hat{\pi}_{j'} - \pi^*_{j'}|_1 \max_{j',j} |\frac{1}{n} \sum_{i=1}^n Z^T_{ij'} \eta_{ij}|_\infty |\beta^*|_1 \\ \leq \; & c' |\beta^*|_1 \sqrt{\frac{64e^2 \rho_\eta^2 \log(dp^2)}{n}} \left(\sqrt{k_1} err_e + \sqrt{k_1} err_a + \max_{j=1,\ldots,p} |\pi^*_{j,S^c_{\tau_j}}|_1 \right) \end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 \log(p \vee d)) - 2\exp(-\log(dp^2)) \geq 1 - c_3' \exp(-c_4' \log(p \vee d))$, for some positive universal constants c_3' and c_4'. Note that, under Assumption 2.5(iii), for some sufficiently large positive universal constant c_0', the bound above is dominated by

$$c_0' |\beta^*|_1 \max_j \sqrt{\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X^{*2}_{ij} \right)} \mathcal{T}_1.$$

Finally, for the term $|\frac{1}{n} X^{*T} \epsilon|_\infty$, we apply Lemma B.1 and a union bound to obtain

$$|\frac{1}{n} X^{*T} \epsilon|_\infty \leq 8e \rho_{X^*} \rho_\epsilon \sqrt{\frac{\log p}{n}}$$

with probability at least $1 - 2\exp(-\log p)$, where we have used the fact that $\mathbb{E}(\frac{1}{n} Z_j^T \epsilon) = \mathbf{0}$ for all j, and the condition $\sqrt{\frac{\log p}{n}} \leq \frac{1}{2}$ implied by Assumption 2.5(i) (so that the term E_1 is no greater than the term E_2 in bound (57)).

Putting everything together, the claim in Lemma A.4 follows. □

A.2 Proof for Theorem 2.2

Under the conditions in Theorem 2.2, $\mathcal{T}_0 \asymp (|\beta^*|_1 \vee 1) \sqrt{\frac{(k_1 \vee 1) \log(d \vee p)}{n}}$ and

$$\lambda_n^{(0)} = \hat{\mathcal{T}}_0^{(0)} \asymp \mathcal{T}_0^{(0)} = \left(\frac{n}{\log(d \vee p)} \right)^{\frac{1}{4}} \sqrt{k_1 \vee 1} \left(\frac{\log(d \vee p)}{n} \right)^{\frac{1}{2} - \varsigma}$$

with probability $1-o(1)$. Since $|\beta^*|_1 = O\left(\left(\frac{n}{\log(d\vee p)}\right)^{\frac{1}{4}}\right)$, $\frac{\log(d\vee p)}{n} = o(1)$, and $\mathcal{T}_0 \geq 2\left|\frac{\hat{X}^T\xi}{n}\right|_\infty$ with probability $1-o(1)$ (by Lemma A.4), we have $2\left|\frac{\hat{X}^T\xi}{n}\right|_\infty = o_p(\lambda_n^{(0)})$. Consequently, by the proof for Lemma A.1, bounds (22)-(23) hold. If $\sqrt{k_2}\bar{B}^{(1)} + |\beta^*_{S_\tau^c}|_1 = o(1)$, we have $\left|\hat{\beta}^{(1)}\right|_1 = |\beta^*|_1 + o(1)$ with probability $1-o(1)$.

By the construction of $\hat{\mathcal{T}}_0^{(1)}$ in the "Adjusted-Penalization" step which uses $\left|\hat{\beta}^{(1)}\right|_1$ for setting $\lambda_n^{(1)}$, we have

$$\lambda_n^{(1)} = \hat{\mathcal{T}}_0^{(1)} \asymp (|\beta^*|_1 \vee 1)\sqrt{k_1 \vee 1}\left(\frac{\log(d\vee p)}{n}\right)^{\frac{1}{2}-\varsigma}$$

with probability $1-o(1)$. Again, $2\left|\frac{\hat{X}^T\xi}{n}\right|_\infty = o_p(\lambda_n^{(1)})$ so the asymptotic bounds (24)-(25) follow from the proof for Lemma A.1. □

B Technical lemmas

Lemma B.1. Let $(W_i)_{i=1}^n$ and $\left(W_i^{'}\right)_{i=1}^n$ consist of independent components, respectively. Suppose there exist parameters ρ and $\rho^{'}$ such that $\sup_{r\geq 1} r^{-\frac{1}{2}}\left(\mathbb{E}\left|W_i\right|^r\right)^{\frac{1}{r}} \leq \rho$ and $\sup_{r\geq 1} r^{-\frac{1}{2}}\left(\mathbb{E}\left|W_i^{'}\right|^r\right)^{\frac{1}{r}} \leq \rho^{'}$ for all $i = 1, ..., n$. Then

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n\left(W_iW_i^{'}\right) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(W_iW_i^{'}\right)\right]\right| \geq \varepsilon\right] \leq 2\exp\left(-n\left(\underbrace{\frac{\varepsilon^2}{32e^2\rho^2\rho^{'2}}}_{E_1} \wedge \underbrace{\frac{\varepsilon}{8e\rho\rho^{'}}}_{E_2}\right)\right). \tag{57}$$

Proof. Note that by the Cauchy-Schwarz inequality,

$$(2r)^{-1}\left(\mathbb{E}\left|W_iW_i^{'}\right|^r\right)^{\frac{1}{r}} \leq \left[(2r)^{-\frac{1}{2}}\left(\mathbb{E}\left|W_i\right|^{2r}\right)^{\frac{1}{2r}}\right]\left[(2r)^{-\frac{1}{2}}\left(\mathbb{E}\left|W_i\right|^{2r}\right)^{\frac{1}{2r}}\right] \leq \rho\rho^{'}$$

for all $i = 1, ..., n$. Consequently, $r^{-1}\left(\mathbb{E}\left|W_iW_i^{'}\right|^r\right)^{\frac{1}{r}} \leq 2\rho\rho^{'}$. Let $\mathbb{E}\left(W_iW_i^{'}\right) = \mu_i$. Definition 5.13 and the proof for Lemma 5.15 in Vershynin (2012) imply that

$$\mathbb{E}\left[\exp\left(t\left(W_iW_i^{'} - \mu_i\right)\right)\right] \leq \exp\left(\frac{t^2}{2}\left(16e^2\rho^2\rho^{'2}\right)\right) \qquad \text{for } |t| \leq \frac{1}{4e\rho\rho^{'}}.$$

By independence, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{t}{n}\sum_{i=1}^n\left(W_iW_i^{'} - \mu_i\right)\right)\right] &= \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{t}{n}\left(W_iW_i^{'} - \mu_i\right)\right)\right] \\
&\leq \prod_{i=1}^n \exp\left(\frac{t^2}{2}\left(\frac{16e^2\rho^2\rho^{'2}}{n^2}\right)\right) \qquad \text{for } |t| \leq \frac{n}{4e\rho\rho^{'}} \\
&= \exp\left(\frac{t^2}{2}\left(\frac{16e^2\rho^2\rho^{'2}}{n}\right)\right) \qquad \text{for } |t| \leq \frac{n}{4e\rho\rho^{'}}.
\end{aligned}$$

Let $\upsilon := \frac{4e\rho\rho'}{\sqrt{n}}$ and $b = \left(\frac{n}{4e\rho\rho'}\right)^{-1}$. Applying the Markov's inequality to the random variable $\exp\left(\frac{t}{n}\sum_{i=1}^n \left(W_i W_i' - \mu_i\right)\right)$ with $t \geq 0$ yields

$$\begin{aligned}
\mathbb{P}\left(\frac{t}{n}\sum_{i=1}^n \left(W_i W_i' - \mu_i\right) \geq \varepsilon\right) &= \mathbb{P}\left(\exp\left(\frac{t}{n}\sum_{i=1}^n \left(W_i W_i' - \mu_i\right)\right) \geq \exp\left(t\varepsilon\right)\right) \\
&\leq \frac{\mathbb{E}\left[\exp\left(\frac{t}{n}\sum_{i=1}^n \left(W_i W_i' - \mu_i\right)\right)\right]}{\exp\left(t\varepsilon\right)} \\
&\leq \underbrace{\exp\left(-t\varepsilon + \frac{t^2}{2}\upsilon^2\right)}_{f(t;\,\varepsilon)} \qquad \text{for } t \in \left[0,\, b^{-1}\right].
\end{aligned}$$

For each fixed $\varepsilon \geq 0$, we then compute $f^*(\varepsilon) := \inf_{t\in[0,\,b^{-1}]} f(t;\,\varepsilon)$. Note that the unconstrained minimum of the function $f(\cdot;\,\varepsilon)$ corresponds to $t^* = \frac{\varepsilon}{\upsilon^2}$. If $\frac{\varepsilon}{\upsilon^2} \leq b^{-1}$, then the unconstrained optimum coincides with the constrained minimum; as a result, $f^*(\varepsilon) = -\frac{\varepsilon^2}{2\upsilon^2}$. If $\frac{\varepsilon}{\upsilon^2} > b^{-1}$, the constrained minimum is attained at the boundary point $t^\dagger = b^{-1}$ since $f(\cdot;\,\varepsilon)$ is a monotonically decreasing function in $[0,\,t^*)$; as a result, $f^*(\varepsilon) = f(t^\dagger;\,\varepsilon) = -\frac{\varepsilon}{b} + \frac{\upsilon^2}{2b^2} \leq -\frac{\varepsilon}{2b}$ (since $\frac{\upsilon^2}{b} < \varepsilon$). Consequently, we have shown that

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \left(W_i W_i'\right) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left(W_i W_i'\right)\right] \geq \varepsilon\right] \leq \exp\left(-n\left(\frac{\varepsilon^2}{32e^2\rho^2\rho'^2} \wedge \frac{\varepsilon}{8e\rho\rho'}\right)\right).$$

Since similar argument also applies to the the left-sided event $\frac{\sum_{i=1}^n \left(W_i W_i'\right)}{n} - \mathbb{E}\left[\frac{\sum_{i=1}^n \left(W_i W_i'\right)}{n}\right] \leq -\varepsilon$, part (i) in Lemma B.1 follows with an additional factor of 2 in front of the tail probability. □

Lemma B.2: Let $U \in \mathbb{R}^{n\times p_1}$ be a random matrix where each row of U is sampled independently; for any unit vector $a \in \mathbb{R}^{p_1}$ and all $i = 1, ..., n$, $\sup_{r\geq 1} r^{-\frac{1}{2}}\left(\mathbb{E}\left|a^T U_i^T\right|^r\right)^{\frac{1}{r}} \leq \tilde{\rho}_U$ for some parameter $\tilde{\rho}_U$, where U_i is the ith row of U. Let $\Sigma_U = \mathbb{E}\left(\frac{U^T U}{n}\right)$, $\underline{\kappa} = \lambda_{\min}\left(\Sigma_U\right)$, and $\bar{\kappa} = \lambda_{\max}\left(\Sigma_U\right)$. If

$$n \gtrsim \left(\frac{\tilde{\rho}_U^4}{\underline{\kappa}^2} \vee 1\right)\log p_1 \tag{58}$$

then there exist positive universal constants c^* and c_1^* such that

$$v^{0T}\frac{U^T U}{n}v^0 \geq \frac{\underline{\kappa}}{2}|v^0|_2^2 - c^*\underline{\kappa}\left(\frac{\tilde{\rho}_U^4}{\underline{\kappa}^2} \vee 1\right)\frac{\log p_1}{n}|v^0|_1^2, \tag{59}$$

$$v^{0T}\frac{U^T U}{n}v^0 \leq \frac{3\bar{\kappa}}{2}|v^0|_2^2 + c^*\underline{\kappa}\left(\frac{\tilde{\rho}_U^4}{\underline{\kappa}^2} \vee 1\right)\frac{\log p_1}{n}|v^0|_1^2, \tag{60}$$

with probability at least $1 - 2\exp(-\log p_1)$.

Proof. This result is essentially shown in Loh and Wainwright (2012), Lemma 1, which is a consequence of their Lemmas 12, 13 and 15. Below we re-phrase the argument in their Lemma 1 to give the readers more guidance. For $s \geq 1$, let $\mathbb{K}(2s) := \{\Delta \in \mathbb{R}^{p_1} : |\Delta|_2 \leq 1\ |\Delta|_0 \leq s\}$ where $|\Delta|_0$ denotes the number of non-zero components in Δ. In view of bound (75) in Lemma

15 of Loh and Wainwright (2012), for some positive universal constant $\bar{c} \geq \frac{2}{c}$, as long as $n \geq 2\bar{c}\log p_1 \left(\frac{\underline{\kappa}^2}{54^2\tilde{\rho}_U^4} \wedge 1\right)^{-1}$ (i.e., (58)) so that $s := \frac{1}{2\bar{c}}\frac{n}{\log p_1}\left(\frac{\underline{\kappa}^2}{54^2\tilde{\rho}_U^4} \wedge 1\right) \geq 1$, the choice of such an s along with the choice $t = \frac{\underline{\kappa}}{54}$ yield

$$\begin{aligned}\mathbb{P}\left[\sup_{\Delta\in\mathbb{K}(2s)}\left|\frac{|U\Delta|_2^2}{n} - \mathbb{E}\left(\frac{|U\Delta|_2^2}{n}\right)\right| \geq \frac{\underline{\kappa}}{54}\right] &\leq 2\exp\left(-cn\left(\frac{\underline{\kappa}^2}{54^2\tilde{\rho}_U^4} \wedge 1\right) + \frac{n}{\bar{c}}\left(\frac{\underline{\kappa}^2}{54^2\tilde{\rho}_U^4} \wedge 1\right)\right) \\ &\leq 2\exp\left(-\frac{n}{\bar{c}}\left(\frac{\underline{\kappa}^2}{54^2\tilde{\rho}_U^4} \wedge 1\right)\right)\end{aligned}$$

where the second inequality follows from $\bar{c} \geq \frac{2}{c}$. Then we can apply bound (70) in Lemma 13 in Loh and Wainwright (2012) with $s := \frac{1}{2\bar{c}}\frac{n}{\log p_1}\left(\frac{\underline{\kappa}^2}{54^2\tilde{\rho}_U^4} \wedge 1\right)$ to obtain bound (59) where we let $c^*\left(\frac{\tilde{\rho}_U^4}{\underline{\kappa}^2} \vee 1\right) = \bar{c}\left(\frac{54^2\tilde{\rho}_U^4}{\underline{\kappa}^2} \vee 1\right)$. If $\frac{n}{\bar{c}}\left(\frac{\underline{\kappa}^2}{54^2\tilde{\rho}_U^4} \wedge 1\right) \geq \log p_1$ (i.e., (58)), we have

$$2\exp\left(-\frac{n}{\bar{c}}\left(\frac{\underline{\kappa}^2}{54^2\tilde{\rho}_U^4} \wedge 1\right)\right) \leq 2\exp(-\log p_1).$$

Bound (60) follows exactly the same argument with only one difference: instead of using bound (70) in Lemma 13 of Loh and Wainwright (2012), bound (71) is used and $\lambda_{\max}(\Sigma_x)$ is replaced with $\bar{\kappa}$. □

Remark. Note that with $U = X^*$, $\underline{\kappa} = \underline{\kappa}_2$, $\tilde{\rho}_U = \tilde{\rho}_{X^*}$, and $p_1 = p$ in Lemma B.2, the bound in Assumption 2.3 follows from (59). Similarly, with $U = Z_j$, $\underline{\kappa} = \underline{\kappa}_1$, $\tilde{\rho}_U = \tilde{\rho}_Z$, $p_1 = d$, and a union bound, we can show: for all $j = 1, ..., p$ and $v^j \in \mathbb{R}^d$, there exist positive universal constants c^*, c_2^* such that

$$v^{jT}\frac{Z_j^T Z_j}{n}v^j \geq \frac{\underline{\kappa}_1}{2}|v^j|_2^2 - c^*\underline{\kappa}_1\left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2} \vee 1\right)\frac{\log d}{n}|v^j|_1^2 \tag{61}$$

with probability at least

$$1 - 2\exp\left(-\frac{n}{\bar{c}}\left(\frac{\underline{\kappa}_1^2}{54^2\tilde{\rho}_Z^4} \wedge 1\right) + \log p\right) \geq 1 - 2\exp\left(-\log\left(p \vee d\right)\right),$$

where the inequality follows as long as $\frac{n}{\bar{c}}\left(\frac{\underline{\kappa}_1^2}{54^2\tilde{\rho}_Z^4} \wedge 1\right) \geq 2\log(p \vee d)$. Consequently, for (61) to hold with probability at least $1 - 2\exp(-\log(p \vee d))$, condition (58) needs to be replaced with

$$n \gtrsim \left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2} \vee 1\right)\log(p \vee d).$$

Under the same condition, we can also show: for all $j = 1, ..., p$ and $v^j \in \mathbb{R}^d$,

$$v^{jT}\frac{Z_j^T Z_j}{n}v^j \leq \frac{3\bar{\kappa}_1}{2}|v^j|_2^2 + c^*\underline{\kappa}_1\left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2} \vee 1\right)\frac{\log d}{n}|v^j|_1^2 \tag{62}$$

with probability at least $1 - 2\exp(-\log(p \vee d))$. The bounds (61) and (62) will be used in the following lemma.

Lemma B.3. Suppose $\underline{\kappa}_1 := \min_j \lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n}Z_j^T Z_j\right]\right)$ is bounded away from zero and $\bar{\kappa}_1 :=$

$\max_j \lambda_{\max}\left(\mathbb{E}\left[\frac{1}{n}Z_j^T Z_j\right]\right)$ is bounded from above. Let the first-stage regularization parameters $\lambda_{n,j} = 16e\rho_\eta\sqrt{\frac{\log(dp)}{n}}$ in (3) and $\tau_j = \underline{\kappa}_1^{-1}\lambda_{n,j}$ for all $j = 1, ..., p$. Suppose: the parts regarding Z_i and η_i in Assumptions 2.1-2.2 hold; $\sqrt{\frac{\log(dp)}{n}} \leq \frac{1}{2}$; bound (61) holds with probability at least $1 - 2\exp(-\log(p \vee d))$; for all $j = 1, ..., p$, there exists a positive universal constant c_2^* such that

$$|\pi_j^*|_1 \lambda_{n,j}^{-1}\frac{b_1 \log d}{n} \leq c_2^* \qquad \text{where } b_1 = \underline{\kappa}_1\left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2} \vee 1\right). \tag{63}$$

Then, (8) and (9) hold with probability at least $1 - c_1\exp(-c_2\log(p \vee d))$ for some positive universal constants c_1 and c_2. Moreover, suppose there exist positive universal constants c_3^*, c_4^* such that

$$\max_j \underline{\kappa}_1\left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2} \vee 1\right)\frac{\log d}{n}\left|S_{\tau_j}\right| \leq c_3^*\bar{\kappa}_1, \tag{64}$$

$$\max_j \underline{\kappa}_1\left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2} \vee 1\right)\frac{\log d}{n}|\pi^*_{j,S^c_{\tau_j}}|_1^2 \leq c_4^*\bar{\kappa}_1\left(err_e + err_a\right)^2, \tag{65}$$

where $err_e = \frac{16e\rho_\eta}{\underline{\kappa}_1}\sqrt{\frac{k_1\log(dp)}{n}}$ and $err_a = \max_{j=1,...,p}|\pi^*_{j,S^c_{\tau_j}}|_1^{\frac{1}{2}}\left(\frac{16e\rho_\eta}{\underline{\kappa}_1}\sqrt{\frac{\log(dp)}{n}}\right)^{\frac{1}{2}}$. If bound (62) holds with probability at least $1 - 2\exp(-\log(p \vee d))$, then (10) holds with probability at least $1 - c_1\exp(-c_2\log(p \vee d))$.

Remark. For the special case $p = 1$ (that is, applying the Lasso to a single equation), the choice of $\lambda_{n,j}$ in Lemma B.3 is more conservative (in terms of universal constants) than the one in Bickel, et. al (2009, Theorem 7.2) which would give any $\lambda_{n,j} > 2\sqrt{2}\rho_\eta\sqrt{\frac{\log d}{n}}$. On the other hand, our assumptions here are more general than those in Bickel, et. al (2009). In particular, Bickel, et. al (2009) assume i.i.d. Gaussian noise, η_{ij}, with zero mean, and fixed Z_j (where the diagonal elements of $\frac{1}{n}Z_j^T Z_j$ are normalized to 1) so they have $\mathbb{E}(\eta_{ij}|Z_{ij}) = 0$; in our case, we allow sub-Gaussian noise and sub-Gaussian designs while only requiring $\mathbb{E}\left(\frac{1}{n}Z_{jl}^T\eta_j\right) = 0$ for all $l = 1, ..., d$. Also note that in the context where $p \geq n$, the first-stage regularization parameters $\lambda_{n,j}$s should have the scaling $\sqrt{\frac{\rho_\eta^2\log(p \vee d)}{n}}$ (instead of $\sqrt{\rho_\eta^2\frac{\log d}{n}}$ for the Lasso estimation in a single equation problem) to take into account the fact that there are p endogenous regressors in the main equation and hence, p regressions to perform simultaneously in the first stage.

Proof. Applying Lemma B.1 with $\varepsilon = 8e\sqrt{\frac{\rho_\eta^2\log(dp)}{n}}$ and a union bound yields

$$\max_j |\frac{1}{n}\sum_{i=1}^n Z_{ij}^T\eta_{ij}|_\infty \leq 8e\sqrt{\frac{\rho_\eta^2\log(dp)}{n}} \tag{66}$$

with probability at least $1 - 2dp\exp(-2\log(dp)) = 1 - 2\exp(-\log(dp))$, where we have used the fact that $\mathbb{E}(\frac{1}{n}Z_j^T\eta_j) = \mathbf{0}$ for all j, and the condition $\sqrt{\frac{\log(dp)}{n}} \leq \frac{1}{2}$ (so that the term E_1 is no greater than the term E_2 in bound (57)). With the choice

$$\lambda_{n,j} = 16e\rho_\eta\sqrt{\frac{\log(dp)}{n}} \geq \max_j |\frac{2}{n}\sum_{i=1}^n Z_{ij}^T\eta_{ij}|_\infty$$

in (3) and $\tau_j = \underline{\kappa}_1^{-1}\lambda_{n,j}$ for all $j = 1, ..., p$, we can then follow the same argument used to show Lemma A.1, where ξ is replaced by η_j, $\hat{X}$ is replaced by Z_j, Y is replaced by X_j, β^* is replaced by π_j^*, $\hat{\beta}_{H2SLS}$ is replaced by $\hat{\pi}_j$, (42) is replaced by (63), and Assumption 2.3 is replaced by (61). This process gives us (8) with probability at least $1 - 2\exp(-\log(dp)) - 2\exp(-\log(p \vee d)) \geq 1 - c_1\exp(-c_2\log(p \vee d))$, for some positive universal constants c_1 and c_2.

Similar to (45), we have

$$|\hat{v}^j|_1 \leq 4\sqrt{\left|S_{\tau_j}\right|}|\hat{v}^j|_2 + 4|\pi^*_{j,S^c_{\tau_j}}|_1, \tag{67}$$

where $\hat{v}^j = \hat{\pi}_j - \pi_j^*$ for $j = 1, ..., p$. Consequently, (9) holds with probability at least $1 - c_1\exp(-c_2\log(p \vee d))$.

Applying the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$ to (67), the bound (62) together with (8) and (9) imply that,

$$\begin{aligned}
\frac{\left|Z_j\hat{v}^j\right|_2^2}{n} &\leq \frac{3\bar{\kappa}_1}{2}\left|\hat{v}^j\right|_2^2 + c^*\underline{\kappa}_1\left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2} \vee 1\right)\frac{\log d}{n}\left|\hat{v}^j\right|_1^2 \\
&\leq \frac{3\bar{\kappa}_1}{2}\left|\hat{v}^j\right|_2^2 + c^{'}\underline{\kappa}_1\left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2} \vee 1\right)\frac{\log d}{n}\left|S_{\tau_j}\right||\hat{v}^j|_2^2 + c^{'}\underline{\kappa}_1\left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2} \vee 1\right)\frac{\log d}{n}|\pi^*_{j,S^c_{\tau_j}}|_1^2 \\
&\leq c^{''}\bar{\kappa}_1\left(err_e + err_a\right)^2
\end{aligned} \tag{68}$$

with probability at least $1 - c_1\exp(-c_2\log(p \vee d))$, where the last inequality follows from conditions (64)-(65). □

Lemma B.4. Suppose: (i) Assumption 2.4 and the parts regarding Z_i and η_i in Assumptions 2.1-2.2 hold; (ii) $\underline{\kappa}_1^{-1}$, $\bar{\kappa}_1$, $\rho_\eta = O(1)$ and (18)-(19) hold; (iii) $\min_{\mathtt{s}\in\mathfrak{M}_j}\kappa_j(\mathtt{s}) = O_p(1)$ for all $j = 1, ..., p$, where

$$\begin{aligned}
\kappa_j(l) &:= \sup_{\Delta\in\mathfrak{S}_j(l)}\frac{1}{n}\Delta^T Z_j^T Z_j\Delta, \\
\mathfrak{S}_j(l) &:= \left\{\Delta \in \mathbb{R}^d\backslash\{\mathbf{0}\} : \left|\Delta_{S^c_{\tau_j}}\right|_0 \leq l,\ |\Delta|_2 = 1\right\}, \\
\mathfrak{M}_j &:= \left\{\mathtt{s} \in \mathbb{N} : \mathtt{s} \leq n,\ \mathtt{s} > c_0^\dagger\left(k_1 \vee 1\right)\kappa_j(\mathtt{s})\right\},
\end{aligned}$$

for some positive universal constant $c_0^\dagger$; (iv) for any $j = 1, ..., p$, if $S_{\tau_j} \neq \varnothing$, $\min_{l\in S_{\tau_j}}|\pi^*_{jl}| > \bar{T}$, where

$$\bar{T} := c_1^\dagger\sqrt{\frac{(k_1 \vee 1)\log(d \vee p)}{n}} \geq \left|\hat{\pi}_j - \pi_j^*\right|_2$$

with probability at least $1 - o(1)$, for some positive universal constant $c_1^\dagger$. If bound (66) holds with probability at least $1 - 2\exp(-\log(dp))$, then $\left(\hat{k}_1 \vee 1\right) \asymp (k_1 \vee 1)$ with probability $1 - o(1)$ as $n \to \infty$, $d \to \infty$, and $p \to \infty$.

Remark. Part (iii) is the so-called "bounded sparse eigenvalue" condition (see, e.g., Bickel, et. al, 2009; Belloni and Chernozhukov, 2013). Lemma 1 in Belloni and Chernozhukov (2013) shows that (iii) holds under a "bounded sparse eigenvalue" condition on the population matrix $\mathbb{E}\left[\frac{1}{n}Z_j^T Z_j\right]$. Along with the assumption $\max_{j=1,...,p}|\pi^*_{j,S^c_{\tau_j}}|_1 = O\left((k_1 \vee 1)\sqrt{\frac{\log d}{n}}\right)$ by (18), condition (iii) ensures that $\hat{k}_1 = O_p\left(k_1 \vee 1\right)$. Condition (iv), known as the "beta-min" condition in the literature,

ensures that $S_{\tau_j} \subseteq J(\hat{\pi}_j)$ for all $j = 1, ..., p$ (and consequently $k_1 \leq \hat{k}_1$) with probability $1 - o(1)$. Note that if $\min_{l \in S_{\tau_j}} |\pi^*_{jl}|$ is bounded away from zero while $\sqrt{\frac{(k_1 \vee 1) \log(d \vee p)}{n}} = o(1)$, then (iv) is satisfied.

Proof. Under conditions (i) and (ii) in Lemma B.4, we have $\max_j \left|\hat{\pi}_j - \pi^*_j\right|_2 = O_p\left(\sqrt{\frac{(k_1 \vee 1) \log(d \vee p)}{n}}\right)$. When $S_{\tau_j} = \varnothing$, clearly $J(\hat{\pi}_j) \supseteq S_{\tau_j}$. Since $|\hat{\pi}_j - \pi^*_j|_\infty \leq |\hat{\pi}_j - \pi^*_j|_2 \leq \bar{T}$ with probability at least $1 - o(1)$, we have that $-\bar{T} + \pi^*_{jl} \leq \hat{\pi}_{jl} \leq \bar{T} + \pi^*_{jl}$ for all l with probability at least $1 - o(1)$. Now when $S_{\tau_j} \neq \varnothing$, under condition (iv), given $\bar{T} < \min_{l \in S_{\tau_j}} |\pi^*_{jl}|$, if $\pi^*_{jl} > 0$ and $l \in S_{\tau_j}$, then the left inequality ensures that $\hat{\pi}_{jl} > 0$ and on the other hand if $\pi^*_{jl} < 0$ and $l \in S_{\tau_j}$, then the right inequality ensures that $\hat{\pi}_{jl} < 0$. In either case, we must have $J(\hat{\pi}_j) \supseteq S_{\tau_j}$ and consequently $\hat{k}_1 \geq k_1$ with probability at least $1 - o(1)$.

To show $\hat{k}_1 = O_p(k_1 \vee 1)$, we modify the proofs of Lemma 2 and Theorem 3 in Belloni and Chernozhukov (2013). The optimality condition of (3) yields that $\frac{1}{n}\sum_{i=1}^n Z_{ijl}\left(X_{ij} - Z_{ij}\hat{\pi}_j\right) = \lambda_{n,j}$ for all $l \in J(\hat{\pi}_j)$ and $j = 1, ..., p$. Let $\hat{\mathbf{s}}_j = \left|J(\hat{\pi}_j) \backslash S_{\tau_j}\right|$. Note that we have

$$\begin{aligned}
\sqrt{|J(\hat{\pi}_j)|}\lambda_{n,j} &\leq \left|\left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^T\left(X_{ij} - Z_{ij}\hat{\pi}_j\right)\right]_{J(\hat{\pi}_j)}\right|_2 \\
&\leq \left|\left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^T\left(X_{ij} - Z_{ij}\pi^*_j\right)\right]_{J(\hat{\pi}_j)}\right|_2 + \left|\left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^T\left(Z_{ij}\pi^*_j - Z_{ij}\hat{\pi}_j\right)\right]_{J(\hat{\pi}_j)}\right|_2.
\end{aligned}$$

Moreover,

$$\left|\left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^T\left(X_{ij} - Z_{ij}\pi^*_j\right)\right]_{J(\hat{\pi}_j)}\right|_2 \leq \sqrt{|J(\hat{\pi}_j)|}\left|\frac{1}{n}Z_j^T\eta_j\right|_\infty$$

and

$$\begin{aligned}
\left|\left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^T\left(Z_{ij}\pi^*_j - Z_{ij}\hat{\pi}_j\right)\right]_{J(\hat{\pi}_j)}\right|_2 &\leq \sup_{\Delta \in \mathfrak{S}_j(\hat{\mathbf{s}}_j)}\left|\frac{1}{n}\Delta^T Z_j^T Z_j\left(\pi^*_j - \hat{\pi}_j\right)\right| \\
&\leq \sup_{\Delta \in \mathfrak{S}_j(\hat{\mathbf{s}}_j)}\sqrt{\frac{1}{n}\Delta^T Z_j^T Z_j \Delta}\sqrt{\frac{1}{n}\sum_{i=1}^n\left(Z_{ij}\hat{\pi}_j - Z_{ij}\pi^*_j\right)^2} \\
&= \sqrt{\kappa_j(\hat{\mathbf{s}}_j)}\sqrt{\frac{1}{n}\sum_{i=1}^n\left(Z_{ij}\hat{\pi}_j - Z_{ij}\pi^*_j\right)^2},
\end{aligned}$$

where we have used the definition of $\kappa_j(\hat{\mathbf{s}}_j)$ and the fact that $\frac{U^TU}{n}$ and $\frac{UU^T}{n}$ have the same maximal eigenvalues. By (66), $\rho_\eta = O(1)$, and the condition (19), there exist positive universal constants c_9, c_{10} such that $c_9\left|\frac{1}{n}Z_j^T\eta_j\right|_\infty \leq c_{10}\sqrt{\frac{\log(d \vee p)}{n}} \leq \lambda_{n,j}$ for all $j = 1, ..., p$, with probability at least $1 - 2\exp(-\log(dp))$; therefore we have

$$\sqrt{|J(\hat{\pi}_j)|}\lambda_{n,j} \leq \frac{1}{c_9}\sqrt{|J(\hat{\pi}_j)|}\lambda_{n,j} + \sqrt{\kappa_j(\hat{\mathbf{s}}_j)}\sqrt{\frac{1}{n}\sum_{i=1}^n\left(Z_{ij}\hat{\pi}_j - Z_{ij}\pi^*_j\right)^2}$$

with probability at least $1-2\exp(-\log(dp))$; furthermore, for some positive universal constant c_{11},

$$\left(1-\frac{1}{c_9}\right)\sqrt{\hat{\mathtt{s}}_j}\lambda_{n,j} \leq \sqrt{\kappa_j(\hat{\mathtt{s}}_j)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Z_{ij}\hat{\pi}_j - Z_{ij}\pi_j^*\right)^2} \leq c_{11}\sqrt{\kappa_j(\hat{\mathtt{s}}_j)\,(k_1\vee 1)}\lambda_{n,j}$$

with probability at least $1-c_1\exp(-c_2\log(p\vee d))$, where in the second inequality, we have used (10) in Assumption 2.4, the conditions $\underline{\kappa}_1^{-1}$, $\bar{\kappa}_1$, $\rho_\eta = O(1)$ and (18), as well as the fact that $\hat{\mathtt{s}}_j \leq |J(\hat{\pi}_j)|$. Consequently, there exists a positive universal constant c_{12} such that

$$\hat{\mathtt{s}}_j \leq c_{12}\kappa_j(\hat{\mathtt{s}}_j)\,(k_1\vee 1) \tag{69}$$

with probability at least $1-c_1\exp(-c_2\log(p\vee d))$.

Let $2c_{12} := c_0^\dagger$. By optimality conditions, $\hat{\mathtt{s}}_j \leq n$. For any $\mathtt{s}\in\mathfrak{M}_j$, suppose $\hat{\mathtt{s}}_j > \mathtt{s}$. By Lemma 3 in Belloni and Chernozhukov (2013), with probability at least $1-c_1\exp(-c_2\log(p\vee d))$, we have $\hat{\mathtt{s}}_j \leq c_{12}\left\lceil\frac{\hat{\mathtt{s}}_j}{\mathtt{s}}\right\rceil\kappa_j(\mathtt{s})\,(k_1\vee 1)$, which further implies that $\mathtt{s}\leq 2c_{12}\kappa_j(\mathtt{s})\,(k_1\vee 1) = c_0^\dagger\kappa_j(\mathtt{s})\,(k_1\vee 1)$ as $\left\lceil\frac{\hat{\mathtt{s}}_j}{\mathtt{s}}\right\rceil \leq 2\frac{\hat{\mathtt{s}}_j}{\mathtt{s}}$. This contradicts that $\mathtt{s}\in\mathfrak{M}_j$. Consequently, we must have $\hat{\mathtt{s}}_j\leq\mathtt{s}$. Another application of (69) with the fact that $\hat{\mathtt{s}}_j\leq\mathtt{s}$ (so $\kappa_j(\hat{\mathtt{s}}_j)\leq\kappa_j(\mathtt{s})$) yields $\hat{\mathtt{s}}_j\leq c_{12}\kappa_j(\mathtt{s})\,(k_1\vee 1)$ with probability at least $1-c_1\exp(-c_2\log(p\vee d))$. Now take the minimum over $\mathtt{s}\in\mathfrak{M}_j$ and since $\min_{\mathtt{s}\in\mathfrak{M}_j}\kappa_j(\mathtt{s}) = O_p(1)$ for all $j=1,...,p$, we have $\hat{k}_1 = O_p\,(k_1\vee 1)$. □

Our next result, Lemma B.5, requires Assumptions B.1 and B.2.

Assumption B.1. For $err_e = \max_j\frac{\sqrt{k_1}\lambda_{n,j}}{\underline{\kappa}_1}$ and $err_a = \max_j|\pi^*_{j,S^c_{\tau_j}}|_1^{\frac{1}{2}}\left(\frac{\lambda_{n,j}}{\underline{\kappa}_1}\right)^{\frac{1}{2}}$,

$$\max_j\underline{\kappa}_1\left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2}\vee 1\right)\frac{\log d}{n}\left|S_{\tau_j}\right| = O\left(\bar{\kappa}_1\right), \tag{70}$$

$$\max_j\underline{\kappa}_1\left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2}\vee 1\right)\frac{\log d}{n}|\pi^*_{j,S^c_{\tau_j}}|_1^2 = O\left(\bar{\kappa}_1\left(err_e+err_a\right)^2\right). \tag{71}$$

Assumption B.2. $\sqrt{k_1}\left(\frac{\log(p\vee d)}{n}\right)^{\frac{1}{4}-\varsigma} = o(1)$ and $\max_j|\pi^*_{j,S^c_{\tau_j}}|_1^{\frac{1}{2}}\left(\frac{\log(p\vee d)}{n}\right)^{\frac{1}{8}-\frac{\varsigma}{2}} = o(1)$.

Lemma B.5. Assume $\eta_{ij}\overset{i.i.d.}{\sim}\mathcal{N}\left(0,\,\sigma_\eta^2\right)$ in (2) for all $j=1,...,p$. Let the first-stage regularization parameters, $\lambda_{n,j}$, in (3) chosen according to (29)-(30), and $\tau_j = \underline{\kappa}_1^{-1}\lambda_{n,j}$ for all $j=1,...,p$. Suppose: Assumptions B.1-B.2 and the parts regarding Z_i, η_i in Assumptions 2.1-2.2 hold; $\underline{\kappa}_1^{-1} = O(1)$, $\bar{\kappa}_1 = O(1)$, $\frac{\log(d\vee p)}{n} = o(1)$ and $\sigma_\eta = o\left(\left(\frac{n}{\log(p\vee d)}\right)^{\frac{1}{4}}\right)$; bounds (61)-(62) hold with probability at least $1-2\exp\left(-\log\left(p\vee d\right)\right)$; for all $j=1,...,p$,

$$|\pi_j^*|_1\lambda_{n,j}^{-1}\frac{b_1\log d}{n} = O\left(1\right) \qquad \text{where } b_1 = \underline{\kappa}_1\left(\frac{\tilde{\rho}_Z^4}{\underline{\kappa}_1^2}\vee 1\right). \tag{72}$$

Then, we have (32) as $n\to\infty$, $d\to\infty$, and $p\to\infty$; moreover, under the conditions (18) and $1 = O\left(\sigma_\eta\right)$, then we also have (33).

Proof. Given $\lambda_{n,j}^{(0)} = \hat{\sigma}_\eta^{(0)}\left(\frac{\log(p\vee d)}{n}\right)^{\frac{1}{2}-\varsigma} = \left(\frac{\log(p\vee d)}{n}\right)^{\frac{1}{4}-\varsigma}$ with $\varsigma\in\left(0,\,\frac{1}{4}\right)$, $\frac{\log(p\vee d)}{n} = o(1)$, and

$\sigma_\eta = o\left(\left(\frac{n}{\log(p\vee d)}\right)^{\frac{1}{4}}\right)$, we have $\max_j |\frac{2}{n}\sum_{i=1}^n Z_{ij}^T\eta_{ij}|_\infty = O_p\left(\sqrt{\frac{\sigma_\eta^2\log(d\vee p)}{n}}\right)$, where the bound follows from (66) and that $\eta_{ij} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_\eta^2\right)$; consequently, $\max_j |\frac{2}{n}\sum_{i=1}^n Z_{ij}^T\eta_{ij}|_\infty = o_p\left(\lambda_{n,j}^{(0)}\right)$. With the choice $\lambda_{n,j} = \lambda_{n,j}^{(0)} = \left(\frac{\log(p\vee d)}{n}\right)^{\frac{1}{4}-\varsigma}$ and $\tau_j = \underline{\kappa}_1^{-1}\lambda_{n,j}$ for all $j = 1, ..., p$, we can then follow the same argument used to show Lemma A.1, where ξ is replaced by η_j, $\hat{X}$ is replaced by Z_j, Y is replaced by X_j, β^* is replaced by π_j^*, $\hat{\beta}_{H2SLS}$ is replaced by $\hat{\pi}_j^{(1)}$, (42) is replaced by (72), and Assumption 2.3 is replaced by (61). This process gives us

$$\left|\pi_j^* - \hat{\pi}_j^{(1)}\right|_2 = O_p\left(err_e + err_a\right), \tag{73}$$

$$\left|\pi_j^* - \hat{\pi}_j^{(1)}\right|_1 = O_p\left(\sqrt{k_1}err_e + \sqrt{k_1}err_a + \max_j |\pi^*_{j,S^c_{\tau_j}}|_1\right). \tag{74}$$

With (8) replaced by (73), (9) replaced by (74), (64) replaced by (70), and (65) replaced by (71), we apply the same argument as what leads to (68) to obtain

$$\begin{aligned}&\max_j \sqrt{\frac{1}{n}\sum_{i=1}^n \left[Z_{ij}\left(\pi_j^* - \hat{\pi}_j^{(1)}\right)\right]^2}\\ =&O_p\left(\sqrt{k_1}\left(\frac{\log(p\vee d)}{n}\right)^{\frac{1}{4}-\varsigma} + \max_j |\pi^*_{j,S^c_{\tau_j}}|_1^{\frac{1}{2}}\left(\frac{\log(p\vee d)}{n}\right)^{\frac{1}{8}-\frac{\varsigma}{2}}\right) = o_p(1)\end{aligned} \tag{75}$$

where we have used Assumption B.2 and the fact that $\underline{\kappa}_1^{-1} = O(1)$, $\bar{\kappa}_1 = O(1)$.

In addition, we apply Lemma B.1 with $\varepsilon = 8e\sqrt{\frac{\sigma_\eta^4\log(p\vee d)}{n}}$ and a union bound yields

$$\max_{j=1,...,p}\left|\frac{1}{n}\sum_{i=1}^n \eta_{ij}^2 - \sigma_\eta^2\right| \leq 8e\sqrt{\frac{\sigma_\eta^4\log(p\vee d)}{n}} \tag{76}$$

with probability at least $1 - 2p\exp(-2\log(p\vee d)) \geq 1 - 2\exp(-\log(p\vee d))$, where we have used the condition $\sqrt{\frac{\log(d\vee p)}{n}} = o(1)$ (so that the term E_1 is no greater than the term E_2 in bound (57)). Consequently, $\max_j\left|\frac{1}{n}\sum_{i=1}^n \eta_{ij}^2 - \sigma_\eta^2\right| = o_p(1)$ since $\sigma_\eta = o\left(\left(\frac{n}{\log(p\vee d)}\right)^{\frac{1}{4}}\right)$.

Putting the pieces together, we obtain

$$\begin{aligned}&\max_j\left|\frac{1}{n}\sum_{i=1}^n\left(X_{ij} - Z_{ij}\hat{\pi}_j^{(1)}\right)^2 - \sigma_{\eta_j}^2\right|\\ \leq\ &\max_j\left(\frac{1}{n}\sum_{i=1}^n\left[Z_{ij}\left(\pi_j^* - \hat{\pi}_j^{(1)}\right)\right]^2 + \left|\frac{2}{n}\sum_{i=1}^n\left[\eta_{ij}Z_{ij}\left(\pi_j^* - \hat{\pi}_j^{(1)}\right)\right]\right| + \left|\frac{1}{n}\sum_{i=1}^n\eta_{ij}^2 - \sigma_{\eta_j}^2\right|\right)\\ \leq\ &\max_j\left(\frac{1}{n}\sum_{i=1}^n\left[Z_{ij}\left(\pi_j^* - \hat{\pi}_j^{(1)}\right)\right]^2 + 2\sqrt{\frac{1}{n}\sum_{i=1}^n\eta_{ij}^2}\sqrt{\frac{1}{n}\sum_{i=1}^n\left[Z_{ij}\left(\pi_j^* - \hat{\pi}_j^{(1)}\right)\right]^2} + \left|\frac{1}{n}\sum_{i=1}^n\eta_{ij}^2 - \sigma_{\eta_j}^2\right|\right)\\ =\ &o_p(1)\end{aligned}$$

where in the second line we have applied a triangle inequality and in the third line we have applied a Cauchy-Schwarz inequality. A continuous mapping theorem implies $\hat{\sigma}_\eta^{(1)} - \sigma_\eta = o_p(1)$.

For the second claim in Lemma B.5, note that from (30) and the fact $\hat{\sigma}_\eta^{(1)} - \sigma_\eta = o_p(1)$ established above,

$$\lambda_{n,j} = \lambda_{n,j}^{(1)} = \hat{\sigma}_\eta^{(1)} \left(\frac{\log(p \vee d)}{n} \right)^{\frac{1}{2}-\varsigma} = \sigma_\eta \left(\frac{\log(p \vee d)}{n} \right)^{\frac{1}{2}-\varsigma} + o_p\left(\left(\frac{\log(p \vee d)}{n} \right)^{\frac{1}{2}-\varsigma} \right),$$

so again we have $\max_j |\frac{2}{n} \sum_{i=1}^n Z_{ij}^T \eta_{ij}|_\infty = o_p\left(\lambda_{n,j}^{(1)}\right)$. The rest of the proof follows exactly the same argument as what leads to (75): in this case, we have

$$\begin{aligned} &\max_j \sqrt{\frac{1}{n} \sum_{i=1}^n \left[Z_{ij} \left(\pi_j^* - \hat{\pi}_j^{(2)} \right) \right]^2} \\ =&O_p\left(\sqrt{\sigma_\eta^2 k_1} \left(\frac{\log(p \vee d)}{n} \right)^{\frac{1}{2}-\varsigma} + \max_j |\pi^*_{j,S^c_{\tau_j}}|_1^{\frac{1}{2}} \sigma_\eta^{\frac{1}{2}} \left(\frac{\log(p \vee d)}{n} \right)^{\frac{1}{4}-\frac{\varsigma}{2}} \right) \end{aligned}$$

Under the conditions (18) and $1 = O(\sigma_\eta)$, (33) follows from the bound above. □

Our next result, Lemma B.6, requires Assumptions B.3-B.6.

Assumption B.3. In terms of $X \in \mathbb{R}^{n \times p}$, for any unit vector $a \in \mathbb{R}^p$, there exists a parameter $\tilde{\rho}_X$ such that $\sup_{r \geq 1} r^{-\frac{1}{2}} \left(\mathbb{E} \left| a^T X_i^T \right|^r \right)^{\frac{1}{r}} \leq \tilde{\rho}_X$, where X_i is the ith row of X; moreover,

$$v^{0T} \frac{X^T X}{n} v^0 \leq \frac{3\bar{\kappa}_X}{2} |v^0|_2^2 + c^* \underline{\kappa}_X \left(\frac{\tilde{\rho}_X^4}{\underline{\kappa}_X^2} \vee 1 \right) \frac{\log p}{n} |v^0|_1^2 \quad \forall v^0 \in \mathbb{R}^p \tag{77}$$

with probability at least $1 - 2\exp(-\log p)$.

Remark. With $U = X$, $\bar{\kappa} = \bar{\kappa}_X$, $\underline{\kappa} = \underline{\kappa}_X$, $\tilde{\rho}_U = \tilde{\rho}_X$, and $p_1 = p$ in Lemma B.2, the bound in Assumption B.3 follows from (60).

Assumption B.4. For $\underline{\kappa}_X := \lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n} X^T X \right] \right)$ and $\bar{\kappa}_X := \lambda_{\max}\left(\mathbb{E}\left[\frac{1}{n} X^T X \right] \right)$,

$$\underline{\kappa}_X \left(\frac{\tilde{\rho}_X^4}{\underline{\kappa}_X^2} \vee 1 \right) \frac{\log p}{n} |S_\tau| = O(\bar{\kappa}_X), \tag{78}$$

$$\underline{\kappa}_X \left(\frac{\tilde{\rho}_X^4}{\underline{\kappa}_X^2} \vee 1 \right) \frac{\log p}{n} |\beta^*_{S^c_\tau}|_1^2 = O\left(\bar{\kappa}_X \left(\frac{\sqrt{k_2}\lambda_n}{\underline{\kappa}_2} + |\beta^*_{S^c_\tau}|_1^{\frac{1}{2}} \left(\frac{\lambda_n}{\underline{\kappa}_2} \right)^{\frac{1}{2}} \right)^2 \right). \tag{79}$$

Assumption B.5. $\frac{\sqrt{k_2}\mathcal{T}_0^{(0)}}{\underline{\kappa}_2} = o(1)$ and $\sqrt{\frac{|\beta^*_{S^c_\tau}|_1 \mathcal{T}_0^{(0)}}{\underline{\kappa}_2}} = o(1)$ where

$$\begin{aligned} \mathcal{T}_0^{(0)} &= \left(\frac{n}{\log(d \vee p)} \right)^{\frac{1}{4}} \max\left\{ \mathcal{T}_{1f}, \sigma_\eta \left(\frac{\log p}{n} \right)^{\frac{1}{2}-\varsigma}, \left(\frac{\log p}{n} \right)^{\frac{1}{2}-\varsigma} \right\}, \\ \mathcal{T}_{1f} &= \left(\sqrt{\sigma_\eta^2 k_1} \vee \sqrt{\sigma_\eta} \right) \left(\frac{\log(p \vee d)}{n} \right)^{\frac{1}{2}-\varsigma}. \end{aligned}$$

Assumption B.6. $\frac{k_2\mathcal{T}_0^{(0)}}{\underline{\kappa}_2} = o(1)$, $\sqrt{\frac{k_2|\beta^*_{S_\tau^c}|_1\mathcal{T}_0^{(0)}}{\underline{\kappa}_2}} = o(1)$, and $|\beta^*_{S_\tau^c}|_1 = o(1)$.

Lemma B.6. Assume $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ in (1) and the same conditions in Lemma B.5 as well as 2.1-2.3, 2.5-2.6, B.3-B.5. Let the second-stage regularization parameter, λ_n, in (4) chosen according to (26)-(28), and $\tau = \underline{\kappa}_2^{-1}\lambda_n$. Suppose

$$|\beta^*|_1\lambda_n^{-1}\left(\frac{b_0\log p}{n} \vee \left(\mathcal{T}_{1f}^2\right)\right) = O(1) \qquad \text{where } b_0 = \underline{\kappa}_2\left(\frac{\tilde{\rho}^4_{X^*}}{\underline{\kappa}_2^2} \vee 1\right). \tag{80}$$

If $\bar{\kappa}_X = O(1)$, $\rho_{X^*} = O(1)$, $\sigma_\epsilon = o\left(\left(\frac{n}{\log(p\vee d)}\right)^{\frac{1}{4}}\right)$, and $|\beta^*|_1 = O\left(\left(\frac{n}{\log(d\vee p)}\right)^{\frac{1}{4}}\right)$, then we have (34) as $n\to\infty$, $d\to\infty$, and $p\to\infty$; moreover, if Assumption B.6 holds and $1 = O\left((\sigma_\epsilon \wedge |\beta^*|_1)\right)$, then (37)-(38) hold.

Proof. Given the form of $\lambda_n^{(0)}$ in (26) and $\mathcal{T}_0 \geq 2\left|\frac{\hat{X}^T\xi}{n}\right|_\infty$ with probability $1-o(1)$ (by Lemma A.4), if Assumption 2.6 and the conditions in Lemma B.5 hold, $\sigma_\epsilon = o\left(\left(\frac{n}{\log(p\vee d)}\right)^{\frac{1}{4}}\right)$, $|\beta^*|_1 = O\left(\left(\frac{n}{\log(d\vee p)}\right)^{\frac{1}{4}}\right)$, and $\rho_{X^*} = O(1)$, we have $2\left|\frac{\hat{X}^T\xi}{n}\right|_\infty = o_p(\lambda_n^{(0)})$ as

$$\begin{aligned}\lambda_n^{(0)} = &\left(\frac{n}{\log(d\vee p)}\right)^{\frac{1}{4}}\max\left\{\left(\sqrt{\sigma_\eta^2 k_1}\vee\sqrt{\sigma_\eta}\right)\left(\frac{\log(p\vee d)}{n}\right)^{\frac{1}{2}-\varsigma}, \sigma_\eta\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\varsigma}, \left(\frac{\log p}{n}\right)^{\frac{1}{2}-\varsigma}\right\}\\ &+\left(\frac{n}{\log(d\vee p)}\right)^{\frac{1}{4}} o\left(\max\left\{\sqrt{k_1\vee 1}\left(\frac{\log(p\vee d)}{n}\right)^{\frac{1}{2}-\varsigma}, \left(\frac{\log p}{n}\right)^{\frac{1}{2}-\varsigma}\right\}\right)\end{aligned}$$

with probability $1-o(1)$. With the choice $\lambda_n = \lambda_n^{(0)}$ and $\tau = \underline{\kappa}_2^{-1}\lambda_n$, we can then follow the same argument used to show Lemma A.1, where (42) is replaced by (80). This process gives us

$$\left|\hat{\beta}^{(1)} - \beta^*\right|_2 = O_p\left(\bar{B}^{(1)}\right), \tag{81}$$

$$\left|\hat{\beta}^{(1)} - \beta^*\right|_1 = O_p\left(\sqrt{k_2}\bar{B}^{(1)} + |\beta^*_{S_\tau^c}|_1\right), \tag{82}$$

where $\bar{B}^{(1)} := \frac{\sqrt{k_2}}{\underline{\kappa}_2}\mathcal{T}_0^{(0)} + \sqrt{\frac{|\beta^*_{S_\tau^c}|_1\mathcal{T}_0^{(0)}}{\underline{\kappa}_2}}$ and $\hat{\beta}^{(1)}$ are the initial estimates returned by Step 1 of the *modified* algorithm based on $\left|\hat{\beta}^{(0)}\right|_1$. With (8) replaced by (81), (9) replaced by (82), (64) replaced by (78), and (65) replaced by (79), we apply the same argument as what leads to (68) to obtain

$$\sqrt{\frac{1}{n}\sum_{i=1}^n\left[X_i\left(\beta^* - \hat{\beta}^{(1)}\right)\right]^2} = O_p\left(\bar{B}^{(1)}\right) = o(1)$$

where we have used Assumption B.5 and the fact that $\bar{\kappa}_X = O(1)$.

In addition, we apply Lemma B.1 with $\varepsilon = 8e\sqrt{\frac{\sigma_\epsilon^4\log(p\vee d)}{n}}$ to obtain

$$\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i^2 - \sigma_\epsilon^2\right| \leq 8e\sqrt{\frac{\sigma_\epsilon^4\log(p\vee d)}{n}} \tag{83}$$

with probability at least $1-2\exp\left(-2\log\left(p\vee d\right)\right)$, where we have used the condition $\sqrt{\frac{\log(p\vee d)}{n}}=o(1)$ (so that the term E_1 is no greater than the term E_2 in bound (57)). Consequently, $\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2-\sigma_\epsilon^2\right|=o_p(1)$ since $\sigma_\epsilon=o\left(\left(\frac{n}{\log(p\vee d)}\right)^{\frac{1}{4}}\right)$.

Putting the pieces together, we obtain

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-X_i\hat{\beta}^{(1)}\right)^2-\sigma_\epsilon^2\right| \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\left[X_i\left(\beta^*-\hat{\beta}^{(1)}\right)\right]^2+\left|\frac{2}{n}\sum_{i=1}^{n}\left[\epsilon_i X_i\left(\beta^*-\hat{\beta}^{(1)}\right)\right]\right|+\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2-\sigma_\epsilon^2\right| \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\left[X_i\left(\beta^*-\hat{\beta}^{(1)}\right)\right]^2+2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[X_i\left(\beta^*-\hat{\beta}^{(1)}\right)\right]^2}+\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2-\sigma_\epsilon^2\right| \\
=\ & o_p(1)
\end{aligned}$$

where in the second line we have applied a triangle inequality and in the third line we have applied a Cauchy-Schwarz inequality. A continuous mapping theorem implies $\hat{\sigma}_\epsilon^{(1)}-\sigma_\epsilon=o_p(1)$.

Under Assumption B.6, we have $\left|\hat{\beta}^{(1)}\right|_1=|\beta^*|_1+o(1)$ with probability $1-o(1)$. By the construction of $\hat{\mathcal{T}}_0^{(1)}$ in (28) which uses $\left|\hat{\beta}^{(1)}\right|_1$ for setting $\lambda_n^{(1)}$, if $1=O\left(\left(\sigma_\epsilon\wedge|\beta^*|_1\right)\right)$, we have

$$\lambda_n=\lambda_n^{(1)}=\hat{\mathcal{T}}_0^{(1)}\asymp\max\left\{\left(|\beta^*|_1\vee\sigma_\epsilon\right)\mathcal{T}_1^f,\,\sigma_\eta\,|\beta^*|_1\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\varsigma},\,\sigma_\epsilon\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\varsigma}\right\}$$

with probability $1-o(1)$. Again, $2\left|\frac{\hat{X}^T\xi}{n}\right|_\infty=o_p(\lambda_n^{(1)})$ so the asymptotic bounds (37)-(38) follow from the proof for Lemma A.1. □

Lemma B.7. Suppose the assumptions in Lemmas A.2 and A.4 hold. Then we have

$$\frac{1}{n}\max_{l\neq j}\left|r_j^T\left(X_l-\hat{X}_l\right)\right|=O_p\left(\left(\left|\hat{\theta}_j\right|_1\vee 1\right)\max\left\{\sigma_{X^*}\mathcal{T}_1,\,\rho_{X^*}\rho_\eta\sqrt{\frac{\log p}{n}}\right\}\right).$$

Proof. Denote $\hat{\eta}_l:=X_l-\hat{X}_l$. Substituting $r_j=\hat{X}_j-\hat{X}_{-j}\hat{\theta}_j$ into the term $\left|r_j^T\hat{\eta}_l\right|$ and applying elementary inequalities yields

$$\frac{1}{n}\left|\hat{X}_j^T\hat{\eta}_l-\hat{\theta}_j^T\hat{X}_{-j}^T\hat{\eta}_l\right|\leq\frac{1}{n}\underbrace{\left|\hat{X}_j^T\hat{\eta}_l\right|}_{\mathscr{T}^*}+\frac{1}{n}\left|\hat{\theta}_j\right|_1\underbrace{\left|\hat{X}_{-j}^T\hat{\eta}_l\right|_\infty}_{\mathscr{T}^\dagger}$$

where

$$\begin{aligned}
\mathscr{T}^* &\leq \underbrace{\left|X_j^{*T}\eta_l^*\right|}_{\mathscr{T}_1}+\underbrace{\left|\left(\hat{X}_j-X_j^*\right)^T\eta_l^*\right|}_{\mathscr{T}_2}+\underbrace{\left|X_j^{*T}\left(\hat{\eta}_l-\eta_l^*\right)\right|}_{\mathscr{T}_3}+\underbrace{\left|\left(\hat{X}_j-X_j^*\right)^T\left(\hat{\eta}_l-\eta_l^*\right)\right|}_{\mathscr{T}_4}, \\
\mathscr{T}^\dagger &\leq \underbrace{\left|X_{-j}^{*T}\eta_l^*\right|_\infty}_{\mathscr{T}_1'}+\underbrace{\left|\left(\hat{X}_{-j}-X_{-j}^*\right)^T\eta_l^*\right|_\infty}_{\mathscr{T}_2'}+\underbrace{\left|X_{-j}^{*T}\left(\hat{\eta}_l-\eta_l^*\right)\right|_\infty}_{\mathscr{T}_3'}+\underbrace{\left|\left(\hat{X}_{-j}-X_{-j}^*\right)^T\left(\hat{\eta}_l-\eta_l^*\right)\right|_\infty}_{\mathscr{T}_4'}.
\end{aligned}$$

The argument used to bound (55) and (56) can be applied to bound $\frac{1}{n}\mathscr{T}_1$, $\frac{1}{n}\mathscr{T}_1'$, $\frac{1}{n}\mathscr{T}_2$, and $\frac{1}{n}\mathscr{T}_2'$; the argument used to bound (52) and (53) can be applied to bound $\frac{1}{n}\mathscr{T}_3$, $\frac{1}{n}\mathscr{T}_3'$, $\frac{1}{n}\mathscr{T}_4$, and $\frac{1}{n}\mathscr{T}_4'$ using the fact that

$$\begin{aligned}\frac{1}{n}\left|X_{-j}^{*T}\left(\hat{\eta}_l-\eta_l^*\right)\right|_\infty &= \frac{1}{n}\left|X_{-j}^{*T}\left(Z_l\left(\pi_l^*-\hat{\pi}_l\right)\right)\right|_\infty,\\ \frac{1}{n}\left|\left(\hat{X}_{-j}-X_{-j}^*\right)^T\left(\hat{\eta}_l-\eta_l^*\right)\right|_\infty &= \frac{1}{n}\max_{j'\neq j}\left|\left(\pi_{j'}^*-\hat{\pi}_{j'}\right)^T Z_{j'}^T Z_l\left(\pi_l^*-\hat{\pi}_l\right)\right|.\end{aligned}$$

□

References

Amemiya, T. (1974). "The non-linear two-stage least squares estimator." *Journal of Econometrics*, 2, 105-110.

Belloni, A. and V. Chernozhukov (2011a). "L1-penalized quantile regression in high-dimensional sparse models." *The Annals of Statistics*, 39, 82-130.

Belloni, A. and V. Chernozhukov (2011b). "High dimensional sparse econometric models: an introduction", in: Inverse Problems and High Dimensional Estimation, Stats in the Château 2009, Alquier, P., E. Gautier, and G. Stoltz, Eds., *Lecture Notes in Statistics*, 203, 127-162, Springer, Berlin.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). "Sparse models and methods for instrumental regression, with an application to eminent domain." *Econometrica*, 80, 2369-2429.

Belloni, A. and V. Chernozhukov (2013). "Least squares after model selection in high-dimensional sparse models." *Bernoulli*, 19, 521-547.

Bickel, P., J. Y. Ritov, and A. B. Tsybakov (2009). "Simultaneous analysis of Lasso and Dantzig selector." *The Annals of Statistics*, 37, 1705-1732.

Bühlmann, P. and S. A. van de Geer (2011). *Statistics for High-Dimensional Data.* Springer, New-York.

Caner, M. (2009). "LASSO type GMM estimator." *Econometric Theory*, 25, 1-23.

Candès, E. and T. Tao (2007). "The Dantzig selector: statistical estimation when p is much larger than n." *The Annals of Statistics*, 35, 2313-2351.

Carrasco, M. (2012). "A regularization approach to the many instruments problem." *Journal of Econometrics*, 170, 383-398.

Fan, J. and R. Li (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of American Statistical Association*, 96, 1348-1360.

Fan, J. and Y. Liao (2014). "Endogeneity in ultrahigh dimension." *The Annals of Statistics*, 42, 872-917.

Fan, J. and J. Lv (2010). "A selective overview of variable selection in high dimensional feature space." *Statistica Sinica*, 20, 101-148.

Fan, J. and J. Lv (2011). "Non-concave penalized likelihood with NP-dimensionality." *IEEE Transactions on Information Theory*, 57, 5467-5484.

Fan, J., J. Lv, and L. Qi (2011). "Sparse high dimensional models in economics." *Annual Review of Economics*, 3, 291-317.

Garen, J. (1984). "The Returns to schooling: a selectivity bias approach with a continuous choice variable." *Econometrica*, 52, 1199-1218.

Gautier, E. and A. B. Tsybakov (2014). "High-dimensional instrumental variables regression and confidence sets." Manuscript. CREST (ENSAE).

Javanmard, A. and A. Montanari (2014). "Confidence intervals and hypothesis testing for high-dimensional regression" *Journal of Machine Learning Research*, 15, 2869-2909.

Jing, B.-Y., Q. M. Shao, and Q. Wang (2003). "Self-normalized Cramér-Type large deviations for independent random variables." *The Annals of Probability*, 31, 2167-2215.

Leeb, H. and Pötscher, B. M. (2006). "Can one estimate the conditional distribution of post-model-selection estimators?" *The Annals of Statistics*, 34, 2554-2591.

Lim, C. and B. Yu. (2013). "Estimation stability with cross validation (ESCV)." arXiv:1303.3128.

Loh, P., and M. Wainwright (2012). "High-dimensional regression with noisy and missing data: provable guarantees with non-convexity." *The Annals of Statistics*, 40, 1637-1664.

Manresa, E. (2015). "Estimating the structure of social interactions using panel data." Working paper. CEMFI.

Meinshausen, N., and P. Bühlmann (2006). "High-dimensional graphs and variable selection with the Lasso." *The Annals of Statistics*, 34:1436-1462.

Meinshausen, N., and B. Yu (2009). "Lasso-type recovery of sparse representations for high-dimensional data." *The Annals of Statistics*, 37, 246-270.

Minsker, S. (2014). "Geometric median and robust estimation in Banach spaces." arXiv:1308.1334v5.

Negahban, S., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers." *Statistical Science*, 27, 538-557.

Nemirovski, A., and D. Yudin (1983). *Problem complexity and method efficiency in optimization.* John Wiley and Sons Inc.

Ravikumar, P., H. Liu, J. Lafferty, and L. Wasserman (2009). "Sparse additive models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 1009-1030.

Ravikumar, P., M. J. Wainwright, and J. Lafferty (2010). "High-dimensional Ising model selection using l_1-regularized Logistic regression." *The Annals of Statistics*, 38, 1287-1319.

Raskutti, G., M. J. Wainwright, and B. Yu (2010). "Restricted eigenvalue conditions for correlated Gaussian designs." *Journal of Machine Learning Research*, 11, 2241-2259.

Raskutti, G., M. J. Wainwright, and B. Yu (2011). "Minimax rates of estimation for high-dimensional linear regression over l_q-balls." *IEEE Trans. Information Theory*, 57, 6976-6994.

Rosenbaum, M. and A. B. Tsybakov (2010). "Sparse recovery under matrix uncertainty." *The Annals of Statistics*, 38, 2620-2651.

Rosenbaum, M. and A. B. Tsybakov (2013). "Improved matrix uncertainty selector", in: *From Probability to Statistics and Back: High-Dimensional Models and Processes* - A Festschrift in Honor of Jon A. Wellner, Banerjee, M. et al. Eds, *IMS Collections*, 9, 276-290, Institute of Mathematical Statistics.

Tibshirani, R. (1996). "Regression shrinkage and selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58, 267-288.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models." *The Annals of Statistics*, 42, 1166-1202.

Vershynin, R. (2012). "Introduction to the non-asymptotic analysis of random matrices", in Eldar, Y. and G. Kutyniok, Eds, *Compressed Sensing: Theory and Applications*, 210-268, Cambridge.

Wainwright, J. M. (2009). "Sharp thresholds for high-dimensional and noisy sparsity recovery Using l_1-constrained quadratic programming (Lasso)." *IEEE Trans. Information Theory*, 55, 2183-2202.

Wainwright, J. M. (2015). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* In prepa-

ration. University of California, Berkeley.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* MIT Press, Cambridge.

Zhang C.-H. and S. S. Zhang (2014). "Confidence intervals for low dimensional parameters in high dimensional linear models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 217-242.

Zhao, P., and B. Yu. (2006). "On model selection consistency of Lasso." *Journal of Machine Learning Research*, 7, 2541-2567.

Zhu, Y. (2013). "Consistent variable selection of the l_1- regularized 2SLS with high-dimensional endogenous regressors and instruments" Manuscript. University of California, Berkeley. (https://sites.google.com/site/yingzhu1215/home/HD2SLS_2013.pdf)